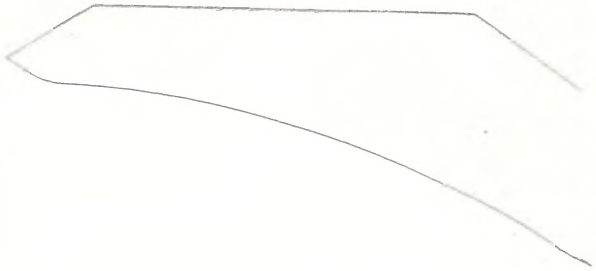
621.391
Т32

Теория передачи сигналов

621.391
Т32

Теория передачи сигналов

Теория передачи сигналов



Теория передачи сигналов

Допущено
Министерством высшего и среднего специального образования СССР
в качестве учебника для студентов, обучающихся по специальностям «Автоматическая электросвязь», «Радиосвязь и радиовещание» и «Многоканальная электросвязь»

МОСКВА «СВЯЗЬ» 1980

ББК 32.81
Т33
УДК 621.391

АВТОРЫ: А. Г. ЗЮКО, Д. Д. КЛОВСКИЙ, М. В. НАЗАРОВ, Л. М. ФИНК

Теория передачи сигналов: Учебник для вузов/Зюко А. Г., Кловский Д. Д., Назаров М. В., Финк Л. М.— М.: Связь, 1980. — 288 с., ил.

Т33

В пер.: 95 к.

Излагаются общие закономерности передачи информации по каналам связи, определяются потенциальные возможности различных способов передачи и приема сигналов, сравниваются различные системы связи между собой и обсуждаются основные направления технической реализации современных систем и перспективы их развития.

Теория передачи сигналов рассматривается как единая научная дисциплина, основу которой составляют теория сигналов, теория помехоустойчивости и теория информации.

Книга предназначена для студентов электротехнических институтов связи. Она также может быть полезной для широкого круга специалистов радиотехники и электросвязи.

$\text{Т}\frac{30401-011}{045(01)-80}\ 5-80$ 2402020000 **ББК 32.81**
6Ф0.1

РЕЦЕНЗЕНТЫ: Н. И. КЛЮЕВ, А. А. КОЛЕСНИКОВ, Н. Т. ПЕТРОВИЧ

Андрей Глебович Зюко, Даниил Давыдович Кловский, Михаил Васильевич Назаров, Лев Матвеевич Финк

ТЕОРИЯ ПЕРЕДАЧИ СИГНАЛОВ

Ответственный редактор Л. М. Финк

Редактор Л. И. Венгренюк

Художник В. Е. Самохин

Художественный редактор Л. А. Даниилин

Технический редактор Г. И. Колосова

Корректор Т. С. Власкина

ИБ № 379

Сдано в набор 28.8.79 г. Подп. в печ. 18.12.79 г.
Т-22624 Формат 60×90/16 Бумага тип. № 2 Гарнитура литературная
Печать высокая Усл. печ. л. 18,0 Уч.-изд. л. 20,33 Тираж 24 000 экз.
Изд. № 18064 Зак. № 168 Цена 95 к.
Издательство «Связь». Москва 101000, Чистопрудный бульвар, д. 2

Типография издательства «Связь» Госкомиздата СССР
Москва 101000, ул. Кирова, д. 40

© Издательство «Связь», 1980 г.

ОСНОВНЫЕ ОБОЗНАЧЕНИЯ

A — случайное сообщение (ансамбль сообщений);
a — реализация сообщения;
B; $B(t)$ — случайный первичный сигнал (сообщение в электрической форме либо последовательность кодовых символов);
$B(t_1, t_2)$; $B(\tau)$ — корреляционная функция;
b, $b(t)$ — реализация первичного сигнала,
C — пропускная способность канала;
$D(\ )$ — дисперсия случайной величины или процесса;
d — расстояние; кодовое расстояние;
E — энергия принимаемого сигнала;
$E_п$ — энергия сигнала на входе канала (на передатчике);
f — частота;
$G(f)$ — спектральная плотность мощности;
$g(t)$; $g(t, \tau)$ — импульсная реакция цепи;
g; g' — выигрыш и обобщенный выигрыш системы модуляции;
H — энтропия дискретной случайной величины (дискретного источника);
H' — производительность источника;
H_ε — эпсилон-энтропия непрерывного источника;
h — дифференциальная энтропия;
h^2 — отношение энергии элемента сигнала на входе демодулятора к спектральной плотности белого шума;
$I(X, Y)$ — взаимная информация между X и Y (информация, содержащаяся в Y относительно X);
$I'(X, Y)$ — скорость передачи информации от X к Y;
K — объем алфавита дискретного источника;
K; $K(i\omega)$ — случайная передаточная функция цепи (канала);
k; $k(i\omega)$ — детерминированная передаточная функция цепи (или реализация случайной передаточной функции);
k — число информационных символов в кодовом блоке;
$M(\ )$ — математическое ожидание случайной величины (процесса);
m — основание кода (объем алфавита кода);
$N(t)$ — случайная помеха (шум);
N_0 — одиосторонняя спектральная плотность мощности белого шума;
$n(t)$ — реализация помехи;
n — длина (число символов) кодовой комбинации (блока);
P — средняя мощность;
$P\{\ \}$; P_x — вероятность события, указанного в скобках или обозначенного индексом;
p — вероятность ошибочного приема символа;
R — скорость передачи информации;
$R(t_1, t_2)$; $R(\tau)$ — нормированная корреляционная функция (коэффициент корреляции);
$S(t)$ — случайный сигнал на выходе цепи (канала) без учета аддитивных помех;
$S(i\omega)$; $S(i2\pi f)$ — комплексная спектральная плотность;
$s(t)$ — реализация сигнала на выходе цепи;
T — длительность тактового интервала; длительность финитного сигнала;
t — текущее время;
$U(t)$ — случайный сигнал на входе канала;
$u(t)$ — реализация случайного сигнала на входе канала;
$V(t)$; $W(t)$; $X(t)$; $Y(t)$ — случайные процессы с реализациями (выборочными функциями) соответственно $v(t)$; $w(t)$; $x(t)$; $y(t)$;
$w(\ )$ — плотность распределения вероятностей;
$Z(t)$ — сумма сигнала и аддитивной помехи на выходе канала;

$z(t)$ — реализация $Z(t)$;
β — индекс угловой модуляции;
$\delta(\ )$ — дельта-функция;
ε; $\varepsilon(t)$ — погрешность оценки случайного параметра или функции;
θ — фазовый сдвиг; текущее время;
$\varkappa$ — избыточность;
Λ — отношение правдоподобия;
λ — информационный параметр;
Π — пик-фактор сообщения или сигнала (отношение максимального значения к среднеквадратичному);
ρ — отношение мощностей сигнала и шума;
σ — среднеквадратичное отклонение;
τ — интервал между двумя сечениями процесса $(t_2 - t_1)$; задержка;
$\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\varphi}$, $\boldsymbol{\psi}$ — фазовый сдвиг; начальная фаза;
$\varphi(t)$; $\psi(t)$ — функции ортогонального (или ортонормированного) базиса;
ω — угловая частота.

Конкретный смысл обозначений уточняется индексами и пояснениями в тексте.

Матрицы и векторы обозначены полужирным шрифтом.

Прямая черта над буквой или формулой означает статистическое усреднение (по ансамблю), волнистая — по времени.

Значок «$\wedge$» над символом означает оценку, выдаваемую демодулятором, декодером или фильтром.

Значок «$\sim$» (тильда) означает преобразование Гильберта.

ПРЕДИСЛОВИЕ

Передача сообщений из одного пункта в другой составляет основную задачу теории и техники связи. В курсе «Теория передачи сигналов» изучаются единые методы решения разнообразных проблем, возникающих при передаче информации от ее источника до получателя.

Жизнь современного общества немыслима без широко разветвленных систем передачи информации. Без них не смогли бы функционировать промышленность, сельское хозяйство, транспорт. Особенно важное значение проблемы передачи информации приобрели в развитом социалистическом обществе, построенном в Советском Союзе. Дальнейшее развитие народного хозяйства и всех сторон деятельности нашего общества немыслимо без широчайшего внедрения автоматизированных систем управления (АСУ), важнейшей частью которых является система связи для обмена информацией, а также устройства хранения и обработки информации.

Передача, хранение и обработка информации имеют место не только при использовании технических устройств. Обычный разговор представляет собой обмен информацией. Всякая книга является хранилищем информации. Астроном, исследующий спектрограмму звезды, извлекает и обрабатывает информацию о химическом составе и физическом состоянии звезды, содержащуюся в ее излучении. Генетическая информация, «записанная» в структуре хромосом клетки, передается при ее делении и содержит «команды», управляющие программой развития организма. По нервным волокнам передается информация от органов чувств к мозгу и от мозга к исполнительным органам. Все вопросы, связанные с передачей информации в природе и обществе, охватывает статистическая теория связи или теория передачи сигналов.

В создании и развитии теории связи большую роль сыграли советские ученые. Академик А. Н. Колмогоров и член-корреспондент АН СССР А. Я. Хинчин разработали необходимый для нее математический аппарат. Академик В. А. Котельников в работах «О пропускной способности «эфира» и проволоки в электросвязи» (1933 г.) и «Теория потенциальной помехоустойчивости» (1946 г.) сформулировал и доказал теоремы, являющиеся основой теории передачи сигналов. Дальнейшее развитие теория получила в трудах А. А. Харкевича, В. И. Сифорова и многих других советских ученых.

Важную роль в становлении современной теории передачи информации сыграл К. Шеннон (США), опубликовавший в 1948 г. свою основополагающую работу «Математическая теория связи». Из других зарубежных ученых, много сделавших для создания и развития теории связи, следует, в первую очередь, упомянуть

Х. Найквиста, Н. Винера, Д. Миддлтона, К. Хелстрома и других.

Без знания основ теории передачи сигналов невозможны создание новых совершенных систем связи и их эксплуатация. Поэтому ее изучение является неотъемлемой частью теоретической подготовки инженеров связи.

Данная книга является учебником по курсу «Теория передачи сигналов». В его основу положены лекции, читавшиеся авторами на протяжении многих лет. Задачей курса является, в первую очередь, формирование на базе современной статистической теории связи, «технического мировоззрения» будущих инженеров-связистов. Поэтому особое внимание в учебнике уделено основным принципам и методам теории передачи сообщений по каналам связи.

Теория передачи сигналов рассматривается как единая научная дисциплина, основу которой составляют теория сигналов, теория статистических решений и теория информации. В соответствии с программой курса вопросы теории информации не выделены в отдельный раздел, а положены в основу всех изучаемых проблем и в той или иной мере содержатся во всех главах учебника. Авторы считают такое построение курса методически обоснованным.

Предполагается, что студенты обладают необходимыми знаниями, полученными при изучении курсов теории линейных и нелинейных электрических цепей, а также математических дисциплин, в частности теории вероятностей, элементов линейной алгебры и теории матриц. Некоторые сведения из теории случайных процессов и функционального анализа, необходимые для современного подхода к изучению теории сигналов, но не входящие в программу курса высшей математики вузов связи, излагаются в гл. 2.

Авторы выражают признательность профессору Н. Т. Петровичу и доцентам А. А. Колесникову и Н. И. Клюеву за полезные замечания при рецензировании рукописи, а также своим коллегам из КЭИС, ЛЭИС, МЭИС и ОЭИС, с которыми они обсуждали ряд вопросов научного и методического характера.

Отзывы по книге просим направить в издательство «Связь» (101000, Москва, Чистопрудный бульвар, 2).

ГЛАВА ПЕРВАЯ

СИСТЕМЫ СВЯЗИ И СПОСОБЫ ПЕРЕДАЧИ СООБЩЕНИЙ

1.1. СООБЩЕНИЕ И СИГНАЛ

Понятия «информация» и «сообщение» употребляют довольно часто. Эти близкие по смыслу понятия сложны, и дать их точное определение через более простые нелегко. В общем случае под *информацией* понимают совокупность сведений о каких-либо событиях, явлениях или предметах. Для передачи или хранения информации используются различные знаки (символы), позволяющие выразить (представить) информацию в некоторой форме. Этими знаками могут быть слова и фразы в человеческой речи, жесты и рисунки, форма колебаний, математические знаки и т. п. Совокупность знаков, содержащих ту или иную информацию, называют *сообщением*. Так, при телеграфной передаче сообщением является текст телеграммы, представляющий собой последовательность отдельных знаков — букв и цифр. При разговоре по телефону сообщением является непрерывное изменение во времени звукового давления, отображающее не только содержание, но и интонацию, тембр, ритм и иные свойства речи. При передаче движущихся изображений в телевизионных системах сообщение представляет собой изменение во времени яркости элементов изображения.

Передача сообщений (а следовательно, и информации) на расстояние осуществляется с помощью какого-либо материального носителя (бумаги, магнитной ленты и т. п.) или физического процесса (звуковых или электромагнитных волн, тока и т. п.). Физический процесс, отображающий (несущий) передаваемое сообщение, называется *сигналом*.

В качестве сигнала можно использовать любой физический процесс, изменяющийся в соответствии с переносимым сообщением. В современных системах управления и связи чаще всего используются электрические сигналы. Физической величиной, определяющей такой сигнал, является ток или напряжение. Сигналы формируются путем изменения тех или иных параметров физического носителя по закону передаваемых сообщений. Этот процесс (изменения параметров носителя) принято называть модуляцией.

Сообщения могут быть функциями времени, например, речь при передаче телефонных разговоров, температура или давление воздуха при передаче телеметрических данных, спектакль при передаче по телевидению и т. п. В других случаях сообщение не является функцией времени (например, текст телеграммы, неподвижное изображение и т. д.).

Сигнал является функцией времени, даже если сообщение таковым не является. Если сигнал представляет собой функцию $u(t)$, принимающую только определенные дискретные значения u_n (например, 1 и 0), то его называют *дискретным*, или, точнее, дискретным по состояниям. Точно так же и сообщение, принимающее только некоторые определенные значения, называется дискретным. Если же сигнал (или сообщение) может принимать любые значения в некотором интервале, то он называется *непрерывным* по состояниям, или *аналоговым*.

В некоторых случаях сообщение или сигнал задается не на всей оси времени, а только в определенные моменты t_n. Такие сообщения (сигналы) называются дискретными по времени, в отличие от непрерывных по времени, заданных на всей оси t. Например, речь является сообщением непрерывным как по состояниям, так и по времени, а датчик температуры, выдающий ее значения через каждые 5 мин, служит источником сообщений, непрерывных по состояниям, но дискретных по времени. Наглядно проиллюстрированы дискретные и непрерывные сигналы на рис. 1.1.

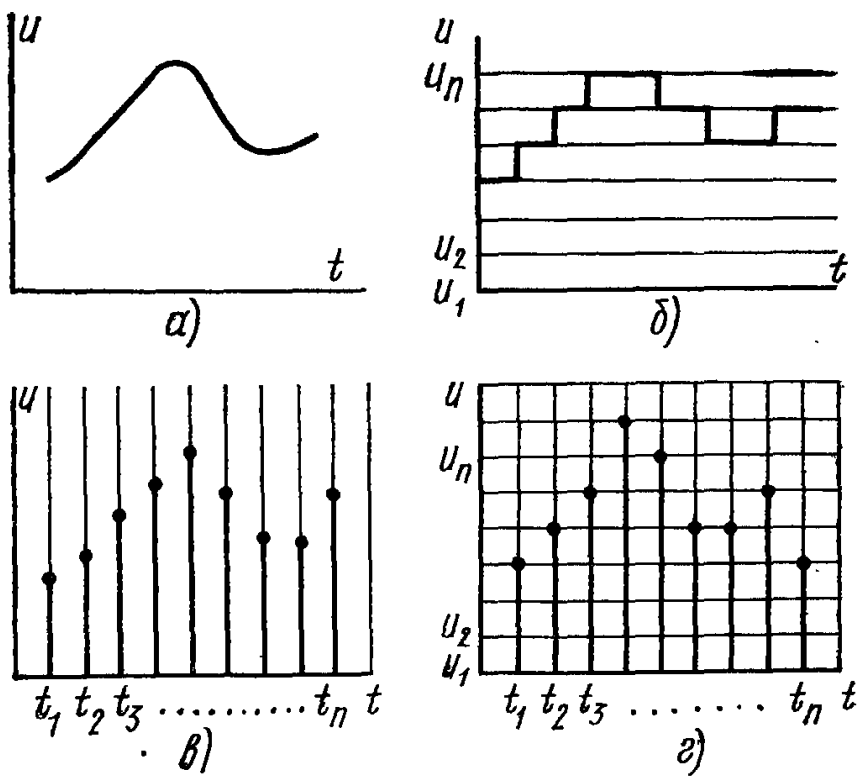


Рис. 1.1. Основные виды сигналов:
а) непрерывный по состояниям и по времени; *б)* дискретный по состояниям, непрерывный по времени; *в)* непрерывный по состояниям, дискретный по времени; *г)* дискретный по состояниям и по времени

Не следует думать, что дискретные сообщения обязательно преобразуются в дискретные сигналы, а непрерывные сообщения — в непрерывные сигналы. В ряде случаев непрерывные сигналы ис-

пользуются для передачи дискретных сообщений, а дискретные сигналы — для передачи непрерывных сообщений.

Сообщение с помощью датчиков обычно преобразуется в электрическую величину $b(t)$ — *первичный сигнал*. При передаче речи такое преобразование выполняет микрофон, при передаче изображения — электронно-лучевая трубка. В большинстве случаев первичный сигнал является низкочастотным колебанием, которое отображает передаваемое сообщение.

В некоторых случаях первичный сигнал непосредственно передается по линии. Так поступают, например, при обычной городской телефонной связи. Для передачи на большие расстояния (по кабелю или радио) первичный сигнал преобразуется в высокочастотный сигнал $u(t)$.

Если бы передаваемое сообщение было детерминированным, т. е. заранее известным с полной достоверностью, то передача его не имела бы никакого смысла. Такое детерминированное сообщение не содержит информации. Поэтому сообщения следует рассматривать как случайные события (или случайные величины, случайные функции). Другими словами, должно существовать некоторое множество вариантов сообщения (например, множество различных результатов хоккейного матча), из которых реализуется с определенной вероятностью одно. Поэтому и сигнал является случайной функцией. Детерминированный сигнал не может быть переносчиком информации. Его можно использовать лишь для испытаний системы связи или отдельных ее элементов.

Случайный характер сообщений, сигналов, а также помех обусловил важнейшее значение математической теории вероятности в построении теории связи. Как будет показано в последующих главах, вероятностные свойства сигналов и сообщений, а также среды, в которой передается сигнал, позволяют определить количество передаваемой информации и ее потери.

Описанием конкретного сигнала может быть некоторая функция времени $s(t)$. Определив так или иначе эту функцию, определяем и сигнал. Однако такое полное описание сигнала не всегда требуется. Для решения ряда вопросов достаточно более общего описания в виде нескольких параметров, характеризующих основные свойства сигнала, подобно тому, как это делается в системах транспортирования. Указывая габариты и массу, мы характеризуем основные свойства предмета с точки зрения условий его транспортировки; другие свойства (например, цвет) с этой точки зрения являются несущественными.

Сигнал также является объектом транспортировки, а техника связи — по существу, техникой транспортирования (передачи) сигналов по каналам связи. Поэтому целесообразно определить параметры сигнала, которые являются основными с точки зрения его передачи. Такими параметрами являются *длительность сигнала* T_c, *его динамический диапазон* H_c *и ширина спектра* F_c.

Всякий сигнал, рассматриваемый как временной процесс, имеет начало и конец. Поэтому длительность сигнала T_c являет-

ся естественным его параметром. Он определяет интервал времени, в пределах которого сигнал существует.

Динамический диапазон определяется как отношение наибольшей мгновенной мощности сигнала к той наименьшей мощности, которую необходимо отличать от нуля при заданном качестве передачи. Он выражается обычно в децибелах. Динамический диапазон речи диктора, например, равен 25—30 дБ, небольшого вокального ансамбля 45—55 дБ, симфонического оркестра 65—75 дБ. Во избежание перегрузок канала в радиовещании динамический диапазон часто сокращают до 35—45 дБ.

Третьим параметром является ширина спектра сигнала F_c. Эта величина дает представление о скорости изменения сигнала внутри интервала его существования. Спектр сигнала, в принципе, может быть неограниченным. Однако для любого сигнала можно указать диапазон частот, в пределах которого сосредоточена его основная энергия. Этим диапазоном и определяется ширина спектра сигнала.

В технике связи спектр сигнала часто сознательно ограничивается. Это обусловлено тем, что аппаратура и линия связи имеют ограниченную полосу пропускаемых частот. Ограничение спектра осуществляется исходя из допустимых искажений сигнала. Например, при телефонной связи требуется, чтобы речь была разборчива и чтобы корреспонденты могли узнать друг друга по голосу. Для выполнения этих условий спектр речевого сигнала можно ограничить полосой от 300 до 3400 Гц. Передача более широкого спектра речи в этом случае нецелесообразна, так как ведет к техническим усложнениям и увеличению затрат. Аналогично необходимая ширина спектра телевизионного сигнала определяется требуемой четкостью изображения. При стандарте в 625 строк, который принят в Советском Союзе, верхняя частота сигнала получается около 6 МГц. Спектр телевизионного сигнала много шире спектра звукового сигнала. Это существенно усложняет построение систем телевизионного вещания по сравнению с системами звукового вещания. Ширина спектра телеграфного сигнала зависит от скорости передачи и обычно принимается равной $F_c \approx 1{,}5v$, где v — скорость телеграфирования в бодах, т. е. количество символов, передаваемых в секунду. Так, при телетайпной передаче $v = 50$ Бод и $F_c \approx 75$ Гц.

Спектр модулированного сигнала обычно шире спектра передаваемого сообщения (первичного сигнала) и зависит от вида модуляции. Можно ввести более общую и наглядную характеристику — *объем сигнала:*

$$V_c = T_c F_c H_c. \qquad (1.1)$$

Объем сигнала V_c дает общее представление о возможностях данного ансамбля сигналов как переносчиков сообщений. Чем больше объем сигнала, тем больше информации можно «вложить» в этот объем и тем труднее передать такой сигнал по каналу связи. Информационный смысл объема сигнала обсуждается в гл. 6.

1.2. СИСТЕМА СВЯЗИ. КАНАЛ СВЯЗИ

На рис. 1.2 изображена структурная схема простейшей системы связи. Рассмотрим назначение отдельных элементов этой схемы. *Источником сообщений* и *получателем* в одних системах связи может быть человек, в других — различного рода устройства (автомат, вычислительная машина и т. п.).

Устройство, преобразующее сообщение в сигнал, называется *передающим устройством*, а устройство, преобразующее принятый сигнал обратно в сообщение, — *приемным устройством*. Передающее устройство включает в себя преобразователь сообщения в первичный сигнал и передатчик. Соответственно приемное устройство состоит из приемника и преобразователя сигнала в сообщение.

С помощью преобразователя в передающем устройстве сообщение a, которое может иметь любую физическую природу (изображение, звуковое колебание и т. п.), преобразуется в первичный электрический сигнал $b(t)$. В телефонии, например, эта операция сводится к преобразованию звукового давления в пропорционально изменяющийся электрический ток микрофона. В телеграфии сначала производится *кодирование*, в результате которого последовательность элементов сообщения (букв) заменяется последовательностью *кодовых* символов (0, 1 или точка, тире), которая затем с помощью телеграфного аппарата преобразуется в последовательность электрических импульсов постоянного тока.

В передатчике первичный сигнал $b(t)$ (обычно низкочастотный) преобразуется во вторичный (высокочастотный) сигнал $u(t)$, пригодный для передачи по используемому каналу. Такое преобразование осуществляется посредством модуляции.

Преобразование сообщения в сигнал должно быть обратимым. В этом случае по выходному сигналу можно, в принципе, восстановить входной первичный сигнал, т. е. получить всю информацию, содержащуюся в переданном сообщении. Если же преобразование необратимо, то часть информации будет потеряна при передаче, даже в тех случаях, когда

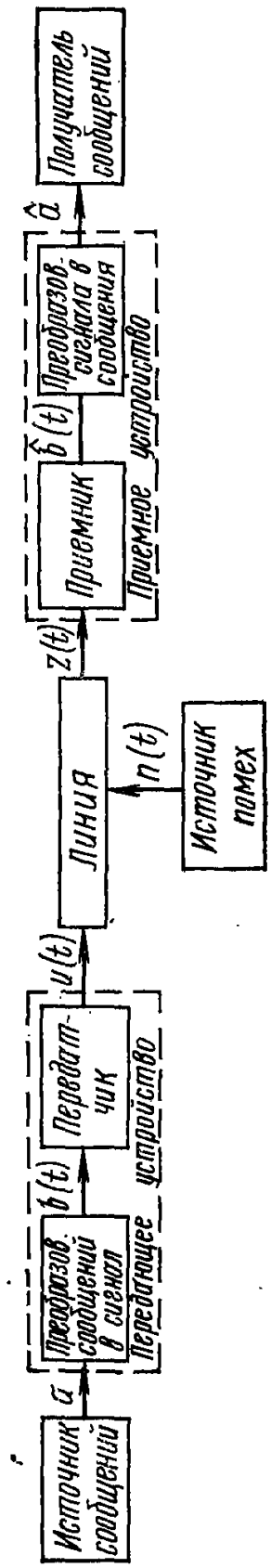


Рис. 1.2. Структурная схема системы связи

сигнал доходит до приемного устройства без искажений.

Линией связи называется среда, используемая для передачи сигналов от передатчика к приемнику. В системах электрической связи — это кабель или волновод, в системах радиосвязи — область пространства, в котором распространяются электромагнитные волны от передатчика к приемнику. При передаче сигнал может искажаться и на него могут накладываться помехи $n(t)$.

Приемное устройство обрабатывает принятое колебание $z(t)=s(t)+n(t)$ и восстанавливает по нему переданное сообщение $\hat{a}$. Другими словами, приемник должен на основе анализа суммарного колебания пришедшего искаженного сигнала $s(t)$, также помехи $n(t)$ определить, какое сообщение a передавалось. Поэтому приемное устройство является одним из наиболее ответственных и сложных элементов системы связи.

Совокупность технических средств, служащих для передачи сообщений от источника к потребителю, называется *системой связи*. Этими средствами являются передающее устройство, линия связи и приемное устройство.

По виду передаваемых сообщений различают следующие системы связи: передачи речи (телефония); передачи текста (телеграфия); передачи неподвижных изображений (фототелеграфия); передачи изображений (телевидение), телеизмерения, телеуправления и передачи данных. По назначению телефонные и телевизионные системы делятся на вещательные, отличающиеся высокой степенью художественности воспроизведения сообщений, и профессиональные, имеющие специальное применение (служебная связь, промышленное телевидение и т. п.).

В системе телеизмерения физическая величина, подлежащая измерению (температура, давление, скорость и т. п.), с помощью датчиков воздействует на передатчик, где она преобразовывается в сигнал и передается по каналу. На приемном конце переданная физическая величина или ее изменения выделяются из сигнала и наблюдаются или регистрируются с помощью приборов.

В системе телеуправления осуществляется передача команд для автоматического выполнения определенных действий. Нередко эти команды формируются автоматически на основании результатов измерения, переданных телеметрической системой.

Системы передачи данных также могут иметь различное применение. В частности, они являются неотъемлемой частью телеметрических и телемеханических систем, автоматизированных систем управления (АСУ).

Каналом связи называется совокупность технических средств, обеспечивающих передачу сигнала от некоторой точки A системы до другой точки B (рис. 1.3). Точки A и B могут быть выбраны произвольно, лишь бы между ними проходил сигнал. Вся часть системы связи, расположенная до точки A, является источником сигнала для этого канала. Если сигналы, поступающие на вход канала и снимаемые с его выхода, являются дискретными

(по состояниям), то канал называется *дискретным*. Если входные и выходные сигналы канала являются непрерывными, то и канал называется *непрерывным*. Встречаются также *дискретно-непрерывные* и *непрерывно-дискретные* каналы, на вход которых поступают дискретные сигналы, а с выхода снимаются непрерывные

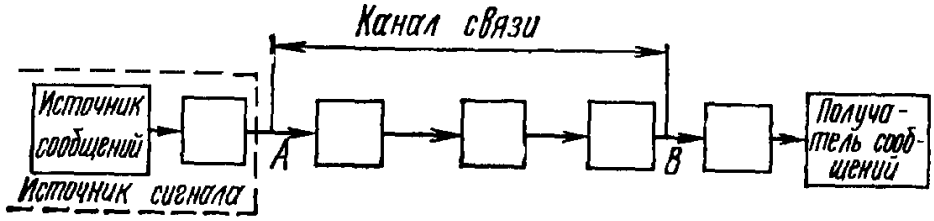


Рис. 1.3. Канал связи

или наоборот. Из сказанного ранее видно, что канал может быть дискретным или непрерывным независимо от характера передаваемых сообщений. Более того, в одной и той же системе связи можно выделить как дискретный, так и непрерывный каналы. Все зависит от того, каким образом выбраны точки A и B входа и выхода канала.

Непрерывный канал связи можно характеризовать так же, как и сигнал, тремя параметрами: временем $T_к$, в течение которого по каналу ведется передача, динамическим диапазоном $H_к$ и полосой пропускания канала $F_к$. Под динамическим диапазоном канала понимают отношение допустимой мощности передаваемого сигнала к мощности неизбежно присутствующей в канале помехи, выраженное в децибелах. Типы каналов, по которым передаются сообщения, многочисленны и разнообразны. Широкое применение находят каналы проводной связи, коротковолновой радиосвязи с использованием отражения от ионосферы, ультракоротковолновой связи ионосферного и тропосферного рассеяния, метеорной связи, космической связи и т. п. Характеристики этих каналов значительно отличаются друг от друга.

Общими признаками непрерывных каналов являются следующие. Во-первых, большинство каналов можно считать линейными. В таких каналах выходной сигнал является просто суммой входных сигналов и помех (применим принцип суперпозиции), а продукты перекрестной модуляции этих сигналов малы по сравнению с выходными сигналами. Во-вторых, на выходе канала даже в отсутствие полезного сигнала всегда имеются помехи. В-третьих, сигнал при передаче по каналу претерпевает задержку по времени и затухание по уровню. И, наконец, в реальных каналах всегда имеют место искажения сигнала, обусловленные несовершенством характеристик канала и, нередко, изменениями параметров канала во времени.

Обобщенной характеристикой непрерывного канала является его емкость (объем):

$$V_к = T_к F_к H_к. \qquad (1.2)$$

Необходимым условием неискаженной передачи по каналу сигналов с объемом V_c, очевидно, должно быть

$$V_c \leqslant V_к. \tag{1.3}$$

Преобразование первичного сигнала в высокочастотный сигнал и преследует цель согласования сигнала с каналом. В простейшем случае сигнал согласуют с каналом по всем трем параметрам, т. е. добиваются выполнения условий:

$$T_c \leqslant T_к;\ F_c \leqslant F_к;\ H_c \leqslant H_к. \tag{1.4}$$

При этих условиях объем сигнала полностью «вписывается» в объем канала. Однако неравенство (1.3) может выполняться и тогда, когда одно или два из неравенств (1.4) не выполнены. Это означает, что можно производить «обмен» длительности на ширину спектра или ширину спектра на динамический диапазон и т. д.

Пусть, например, записанный на пленку телефонный сигнал, имеющий ширину спектра 3 кГц, необходимо передать через канал, полоса пропускания которого равна 300 Гц. Это можно осуществить, воспроизводя сигнал со скоростью, в 10 раз меньшей, чем та, с которой он был записан. При этом все частоты исходного сигнала уменьшатся в 10 раз и во столько же раз увеличится время передачи. Принятый сигнал при этом также записывается на пленку, а затем, воспроизведя его со скоростью, в 10 раз большей, можно восстановить исходный сигнал. Аналогичным образом можно передать сигнал быстрее, чем он создавался, если полоса пропускания канала шире спектра сигнала.

Значительно больший интерес представляет возможность обмена динамического диапазона на полосу пропускания. Так, используя широкополосные помехоустойчивые виды модуляции, которые будут рассматриваться в гл. 6, можно передать сообщение с динамическим диапазоном, скажем, 60 дБ, по каналу, в котором сигнал превышает помеху всего лишь на 20 дБ. При этом используется полоса пропускания канала, в несколько раз более широкая, чем спектр сообщения.

Система связи называется *многоканальной*, если она обеспечивает передачу нескольких сообщений по одной общей линии

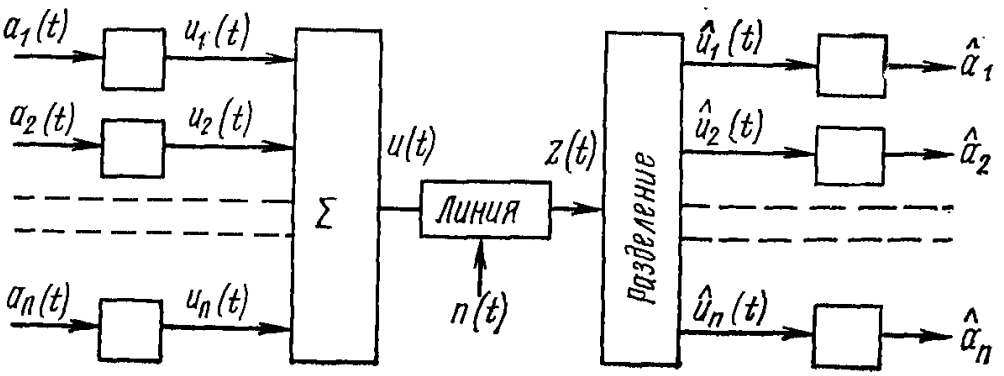


Рис. 1.4. Структурная схема многоканальной системы связи

связи. Структурная схема простейшей многоканальной системы связи изображена на рис. 1.4. Здесь сообщения a_1, a_2, ..., a_n, подлежащие передаче, преобразуются в электрические сигналы $u_1(t)$, $u_2(t)$, ..., $u_n(t)$, а затем смешиваются в аппаратуре уплотнения. Полученный таким образом групповой сигнал $u(t)$ передается по линии связи. Приемник преобразует принятое колебание $z(t)=s(t)+n(t)$ в исходный групповой сигнал, из которого затем с помощью устройства разделения выделяются индивидуальные сигналы $\hat{u}_i(t)$, преобразуемые в соответствующие сообщения $\hat{a}_1$, $\hat{a}_2$, ..., $\hat{a}_n$. Для разделения сигналов на приемном конце, очевидно, необходимо, чтобы они различались между собой по некоторому признаку. В практике многоканальной связи преимущественное применение имеют частотный и временной способы разделения.

1.3. ПОМЕХИ И ИСКАЖЕНИЯ В КАНАЛЕ

В реальном канале сигнал при передаче искажается и сообщение воспроизводится с некоторой ошибкой. Причиной таких ошибок являются искажения, вносимые самим каналом, и помехи, воздействующие на сигнал.

Частотные и временные характеристики канала определяют так называемые линейные искажения. Кроме того, канал может вносить и нелинейные искажения, обусловленные нелинейностью тех или иных звеньев канала. Если линейные и нелинейные искажения обусловлены известными характеристиками канала, то они, по крайней мере в принципе, могут быть устранены путем надлежащей коррекции.

Следует четко отличать искаженя от помех, имеющих случайный характер. Помехи заранее не известны и поэтому не могут быть полностью устранены.

Под *помехой* понимается любое воздействие на полезный сигнал, затрудняющее его прием. Помехи весьма разнообразны как по своему происхождению, так и по физическим свойствам. В радиоканалах наиболее распространеннымн являются *атмосферные помехи*, обусловленные электрическими процессами в атмосфере и, прежде всего, грозовыми разрядами. Энергия этих помех сосредоточена, главным образом, в области длинных и средних волн. Сильные помехи создаются также промышленными установками. Это так называемые *индустриальные помехи*, возникающие из-за резких изменений тока в электрических цепях всевозможных электроустройств. Сюда относятся помехи от электротранспорта, электрических двигателей, медицинских установок, систем зажигания двигателей и т. п.

Распространенным видом помех являются *помехи от посторонних радиостанций* и *каналов*. Они обусловлены нарушением регламента распределения рабочих частот, недостаточной стабильностью частот и плохой фильтрацией гармоник сигнала, а также

нелинейными процессами в каналах, ведущими к перекрестным искажениям.

В проводных каналах связи основным видом помех являются *импульсные* шумы и *прерывания* связи. Появление импульсных помех часто связано с автоматической коммутацией и перекрестными наводками. Прерывание связи есть явление, при котором сигнал в линии резко затухает или совсем исчезает. Такие прерывания могут быть вызваны различными причинами, из которых наиболее частыми являются нарушения контактов в реле.

Практически в любом диапазоне частот имеют место *внутренние шумы* аппаратуры, обусловленные хаотическим движением носителей заряда в усилительных приборах, сопротивлениях и других элементах аппаратуры. Эти помехи особенно сказываются при радиосвязи в диапазоне ультракоротких волн, где другие помехи невелики. В этом диапазоне имеют значение и *космические помехи*, связанные с электромагнитными процессами, происходящими на Солнце, звездах и других внеземных объектах.

В общем виде влияние помехи $n(t)$ на передаваемый сигнал $u(t)$ можно выразить оператором

$$z = \psi(u, n). \tag{1.5}$$

В частном случае, когда оператор ψ вырождается в сумму

$$z = u + n, \tag{1.6}$$

помеха называется *аддитивной*. Если же оператор может быть представлен в виде произведения

$$z = ku, \tag{1.7}$$

то помеху называют *мультипликативной*. Здесь $k(t)$ — случайный процесс[1]. В реальных каналах обычно имеют место и аддитивные и мультипликативные помехи, и поэтому

$$z = ku + n = s + n. \tag{1.8}$$

Среди аддитивных помех различного происхождения особое место занимает *флуктуационная помеха* (флуктуационный шум), представляющая собой случайный процесс с нормальным распределением (гауссовский процесс). Такая помеха наиболее изучена и представляет наибольший интерес как в теоретическом, так и в практическом отношении. Этот вид помех практически имеет место во всех реальных каналах.

С физической точки зрения такие помехи порождаются различного рода флуктуациями, т. е. случайными отклонениями тех или иных физических величин от их средних значений. Так источником шума в электрических цепях могут быть флуктуации тока, обусловленные дискретной природой носителей заряда (электро-

[1] Строгие определения случайного процесса и его энергетического спектра будут даны в гл. 2, а более подробное описание помех — в гл. 3.

нов, ионов). Дискретная природа электрического тока проявляется в электронных лампах и полупроводниковых приборах в виде дробового эффекта.

Сумма большого числа любых помех от различных источников также имеет характер флуктуационной помехи. И, наконец, многие помехи при прохождении через приемное устройство часто приобретают свойства нормальной флуктуационной помехи.

Наиболее распространенной причиной шума являются флуктуации, обусловленные тепловым движением. Случайное тепловое движение носителей заряда в любом проводнике вызывает случайную разность потенциалов (напряжение) на его концах. Среднее значение напряжения равно нулю, а переменная составляющая проявляется как шум. Тепловой шум на входе приемника представляет собой нормальный случайный процесс с нулевым средним и энергетическим спектром:

$$G_{ш}(f)=\frac{hf}{2\left[\exp\left(hf/kT\right)-1\right]}, \tag{1.9}$$

где $h \approx 6{,}6\cdot10^{-34}$ Дж·с — постоянная Планка; $k \approx 1{,}38\cdot10^{-23}$ Дж/град. — постоянная Больцмана; T — абсолютная температура источника шума; f — текущая частота.

В диапазоне звуковых и радиочастот $hf \ll kT$, и поэтому спектральная плотность постоянна и равна

$$G_{ш}(f)=kT/2=N_0/2. \tag{1.10}$$

Величину $N_0=kT$ называют односторонней спектральной плотностью шума. При ширине полосы пропускания приемника F мощность шума равна $P_{ш}=N_0F$, Вт.

В диапазоне оптических частот, который с развитием квантовой электроники становится весьма перспективным для связи, наоборот, $hf \gg kT$ и тепловой шум оказывается очень слабым. Однако в этом диапазоне существенное значение получает «квантовый шум», вызванный дискретной природой излучения сигнала. Сущность квантового шума связана с соотношением неопределенности, согласно которому средние квадратичные ошибки при измерении энергии фотона σ_E и времени его прихода σ_t подчиняются неравенству $\sigma_E\sigma_t \geqslant h$. Поэтому даже при отсутствии аддитивных помех сигнал не может быть принят абсолютно точно. В первом приближении можно рассматривать квантовый шум как помеху со спектральной плотностью, равной энергии фотона hf. В оптическом диапазоне частота f выше 10^{15} Гц, поэтому квантовый шум весьма ощутим.

К *импульсным*, или сосредоточенным по времени, помехам относят помехи в виде одиночных импульсов, следующих один за другим через такие большие промежутки времени, что переходные явления в приемнике от одного импульса успевают практически затухнуть к моменту прихода следующего импульса. К таким помехам относятся многие виды атмосферных и индустриальных помех. Заметим, что понятия «флуктуационная помеха» и «импульсная помеха» являются понятиями относительными. В зависимости от частоты следования импульсов одна и та же помеха может воздействовать как импульсная на приемник с широ-

кой полосой пропускания и как флуктуационная на приемник с относительно узкой полосой пропускания. Импульсные помехи представляют собой случайный процесс, состоящий из отдельных редких, случайно распределенных во времени и по амплитуде, импульсов. Статистические свойства таких помех с достаточной для практических целей полнотой описываются распределением вероятностей амплитуд импульсов и распределением временных интервалов между этими импульсами.

К *сосредоточенным* по спектру помехам принято относить сигналы посторонних радиостанций, излучения генераторов высокой частоты различного назначения (промышленных, медицинских) и т. п. В общем случае это модулированные колебания, т. е. квазигармонические колебания с изменяющимися параметрами. В одних случаях эти колебания являются непрерывными (например, сигналы вещательных и телевизионных радиостанций), в других случаях они носят импульсный характер (сигналы радиотелеграфных станций). В отличие от флуктуационных и импульсных помех, ширина спектра сосредоточенной помехи в большинстве случаев не превышает полосы пропускания приемника. В диапазоне коротких волн этот вид помех является основным, определяющим качество связи.

1.4. КОДИРОВАНИЕ И МОДУЛЯЦИЯ

Преобразование дискретного сообщения в сигнал обычно осуществляется в виде двух операций — кодирования и модуляции.

Кодирование представляет собой преобразование сообщения в последовательность некоторых символов. Для этого устанавливают взаимооднозначное соответствие между сообщениями и символами, которое называется кодом. Код должен быть известен (заложен в аппаратуру) как на передающей, так и на приемной сторонах.

Модуляция представляет собой преобразование сообщения (первичного сигнала) в сигнал, пригодный для передачи по данной линии связи. При этом преобразовании осуществляется согласование источника с каналом.

Простейшим примером дискретного сообщения является текст. Любой текст состоит из конечного числа элементов: букв, цифр, знаков препинания. Их совокупность называется *алфавитом* источника сообщения. Так как число элементов в алфавите конечно, то их можно пронумеровать и тем самым свести передачу сообщения к передаче последовательности чисел.

Так, для передачи букв русского алфавита (их 32) необходимо передать числа от 1 до 32. Для передачи любого числа, записанного в десятичной форме, требуется передача десяти цифр — от 0 до 9. Практически для этого нужно передать по каналу связи десять сигналов, соответствующих различным цифрам. Систему передачи дискретных сообщений можно существенно упрос-

тить, если воспользоваться при кодировании двоичной системой счисления.

В десятичной системе основанием счисления является число 10. Поэтому любое целое число N можно представить в виде

$$N = \ldots a_2 \cdot 10^2 + a_1 \cdot 10^1 + a_0 \cdot 10^0, \qquad (1.11)$$

где a_0, a_1, ..., a_n — коэффициенты, принимающие значения от 0 до 9. Так, число 265 можно записать как $2 \cdot 10^2 + 6 \cdot 10^1 + 5 \cdot 10^0$. Очевидно, в качестве основания счисления можно принять любое целое число m и представить число N как

$$N = \ldots + a_2 m^2 + a_1 m^1 + a_0 m^0, \qquad (1.12)$$

где a_0, a_1, ..., a_n — коэффициенты, принимающие значения от 0 до m—1. Задаваясь величиной m, можно построить любую систему счисления.

При $m=2$ получим двоичную систему, в которой числа записываются с помощью всего лишь двух цифр — 0 и 1. Например, число 13 в двоичной системе записывается 1101, что соответствует выражению $1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$. Арифметические действия в двоичной системе весьма просты. Так, сложение осуществляется по следующим правилам: 0+0=0; 0+1=1; 1+0=1; 1+1=10.

Если преобразовать последовательность элементов сообщения в последовательность двоичных чисел, то для передачи последних по каналу связи достаточно передавать всего лишь два кодовых символа — 0 и 1. Например, символы 0 и 1 могут передаваться колебаниями с различными частотами или импульсами тока разной полярности. Благодаря своей простоте двоичная система счисления нашла широкое применение при кодировании дискретных сообщений.

При кодировании происходит процесс преобразования элементов сообщения в соответствующие им числа (кодовые символы). Каждому элементу сообщения присваивается определенная совокупность кодовых символов, которая называется *кодовой комбинацией*. Совокупность кодовых комбинаций, обозначающих дискретные сообщения, образует *код*. Правило кодирования может быть выражено кодовой таблицей, в которой приводятся алфавит кодируемых сообщений и соответствующие им кодовые комбинации. Множество возможных кодовых символов называется кодовым алфавитом, а их количество m — основанием кода. В общем случае при основании кода m правила кодирования N элементов сообщения сводятся к правилам записи N различных чисел в m-ичной системе счисления. Число разрядов n, образующих кодовую комбинацию, называется значностью кода, или длиной кодовой комбинации. В зависимости от системы счисления, используемой при кодировании, различают *двоичные* и m-ичные (недвоичные) коды.

Различают коды равномерные и неравномерные.

Равномерными называют такие коды, у которых все комбинации имеют одинаковую длину. Для равномерного кода число возможных комбинаций равно m^n. Примером такого кода является пятизначный код Бодо, содержащий пять двоичных элементов ($m=2$, $n=5$). Число возможных кодовых комбинаций равно $2^5=32$, что достаточно для кодирования всех букв алфавита.

Применение равномерных кодов упрощает построение автоматических буквопечатающих устройств и не требует передачи разделительных символов между кодовыми комбинациями.

Неравномерные коды характерны тем, что у них кодовые комбинации отличаются друг от друга не только взаимным расположением символов, но и их количеством. Это приводит к тому, что различные комбинации имеют различную длительность. Типичным примером неравномерных кодов является код Морзе, в котором символы 0 и 1 используются только в двух сочетаниях — как одиночные (1 и 0) или как тройные (111 и 000). Сигнал, соответствующий одной единице, называется точкой, трем единицам — тире. Символ 0 используется как знак, отделяющий точку от тире, точку от точки и тире от тире. Совокупность 000 используется как разделительный знак между кодовыми комбинациями.

По помехоустойчивости коды делятся на *простые* (примитивные) и *корректирующие*. Коды, у которых все возможные кодовые комбинации используются для передачи информации, называются простыми, или кодами без избыточности. В простых равномерных кодах превращение одного символа комбинации в другой, например 1 в 0 или 0 в 1, приводит к появлению новой комбинации, т. е. к ошибке. Корректирующие коды строятся так, что для передачи сообщения используются не все кодовые комбинации, а лишь некоторая часть их. Тем самым создается возможность обнаружения и исправления ошибки при неправильном воспроизведении некоторого числа символов. Корректирующие свойства кодов достигаются ценой введения в кодовые комбинации дополнительных (избыточных) символов (см. ниже, гл. 5).

Декодирование состоит в восстановлении сообщения по принимаемым кодовым символам. Устройства, осуществляющие кодирование и декодирование, называются соответственно *кодером* и *декодером*. Как правило, это логические устройства. На рис. 1.5

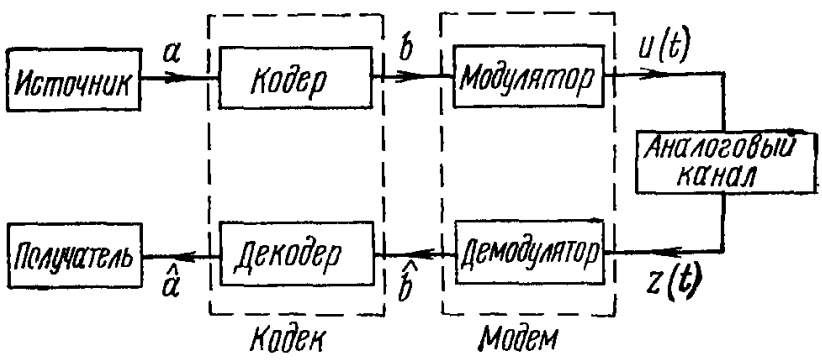


Рис. 1.5. Структурная схема системы передачи дискретных сообщений

изображена структурная схема системы передачи дискретных сообщений, а на рис. 1.6 поясняется процесс преобразования дискретного сообщения в сигнал. Передаваемое сообщение обозна-

чено буквой a, кодированное сообщение (или первичный сигнал) — b, сигнал, поступающий в линию связи, — $u(t)$, принятое колебание — $z(t)$, восстановленная последовательность кодовых символов — $\hat{b}$ и декодированное (восстановленное) сообщение — $\hat{a}$. Обозначения принятых сигналов, кодовых символов и восста-

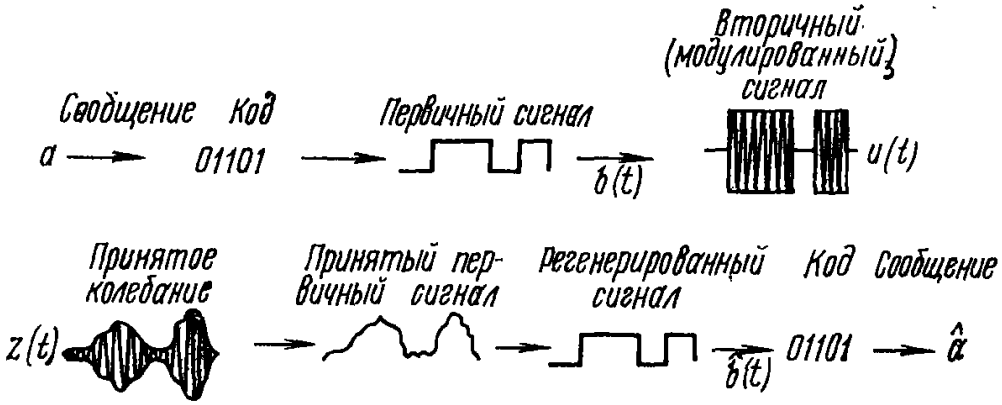


Рис. 1.6. Процесс преобразования дискретного сообщения в сигнал и восстановления переданного сообщения в приемнике

новленного сообщения выбраны иными, чем передаваемых. Этим подчеркивается то обстоятельство, что из-за влияния помех принятый сигнал отличается от переданного, а восстановленное сообщение может не совпадать с исходным.

В современных системах передачи дискретных сообщений принято различать две группы относительно самостоятельных устройств: кодеки и модемы. *Кодеком* называются устройства, преобразующие сообщение в код (кодер) и код в сообщение (декодер), а *модемом* — устройства, преобразующие код в сигнал (модулятор) и сигнал в код (демодулятор).

При передаче непрерывного сообщения a оно сначала преобразуется в первичный электрический сигнал $b(t)$, а затем, как правило, с помощью модулятора формируется сигнал $u(t)$, который и посылается в линию связи. Принятое колебание $z(t)$ подвергается обратным преобразованиям, в результате которых выделяется первичный сигнал $\hat{b}(t)$. По нему затем восстанавливается с той или иной точностью сообщение $\hat{a}$.

Общие принципы модуляции предполагаются известными. Остановимся кратко на особенностях дискретной модуляции.

При дискретной модуляции закодированное сообщение a, представляющее собой последовательность кодовых символов $\{b_i\}$, преобразуется в последовательность элементов (посылок) сигнала $\{u_i\}$. В частном случае дискретная модуляция сводится к воздействию кодовых символов на переносчик $f(t)$.

Посредством модуляции один из параметров переносчика изменяется по закону, определяемому кодом. При непосредственной передаче переносчиком может быть постоянный ток, изменяющимися параметрами которого являются величина и направление тока. Обычно же в качестве переносчика, как и в непрерывной

модуляции, используется переменный ток (гармоническое колебание). В этом случае можно получить амплитудную (АМ), частотную (ЧМ) и фазовую (ФМ) модуляции. Дискретную модуляцию часто называют манипуляцией, а устройство, осуществляющее дискретную модуляцию (дискретный модулятор), называют манипулятором или генератором сигналов.

На рис. 1.7 приведены формы сигналов при двоичном коде для различных видов манипуляции. При АМ символу 1 соответствует передача несущего колебания в течение времени T (посылка), символу 0 — отсутствие колебания (пауза). При ЧМ передача несущего колебания с частотой f_1 соответствует символу 1, а передача колебания с частотой f_0 соответствует 0. При двоичной ФМ меняется фаза несущей на 180° при каждом переходе от 1 к 0 и от 0 к 1.

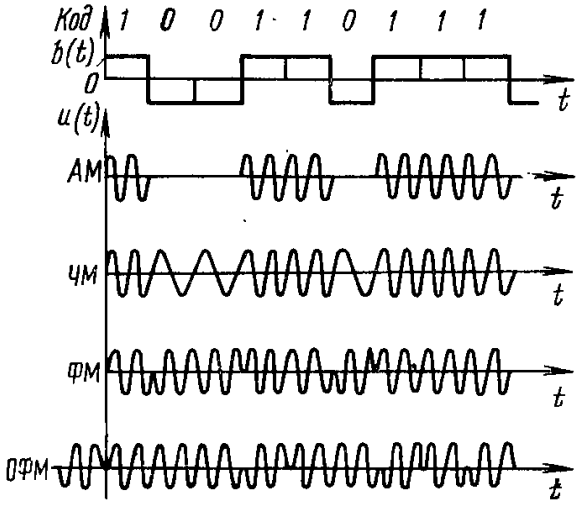


Рис. 1.7. Сигналы при различных видах дискретной модуляции

Наконец, на практике нашла применение система относительной фазовой модуляции (ОФМ). В отличие от ФМ, при ОФМ фаза сигналов отсчитывается не от некоторого эталона, а от фазы предыдущего элемента сигнала. В двоичном случае символ 0 передается отрезком синусоиды с начальной фазой предшествующего элемента сигнала, а символ 1 — таким же отрезком с начальной фазой, отличающейся от начальной фазы предшествующего элемента сигнала на π.

При ОФМ передача начинается с посылки одного, не несущего информации элемента, который служит опорным сигналом для сравнения фазы последующего элемента. Подробнее о приеме таких сигналов и о преимуществах относительного метода модуляции будет сказано в гл. 4.

В более общем случае дискретную модуляцию следует рассматривать как преобразование кодовых символов 0, 1, ..., $m-1$ в определенные отрезки сигнала $u_i(t)$, где $i=0, 1, ..., m-1$ — передаваемый символ. При этом вид отрезка сигнала $u_i(t)$, в принципе, может быть произволен. В действительности его выбирают так, чтобы удовлетворить требованиям, предъявляемым к системе связи (в частности, по скорости передачи и по занимаемой полосе частот), и чтобы сигналы хорошо различались с учетом воздействующих помех.

Длительность посылки первичного сигнала $b(t)$ при дискретной передаче определяет скорость передачи посылок (техническую скорость или скорость телеграфирования). Эта скорость v выражается числом посылок, передаваемых за единицу времени.

Измеряется техническая скорость в бодах. Один Бод — это скорость, при которой за 1 с передается одна посылка. Если длительность посылки T выражена в секундах, то скорость телеграфирования будет равна $v=1/T$, Бод. Частота манипуляции $F_м=1/2T=v/2$, Гц. Если полосу частот ограничить третьей гармоникой, то ширина спектра первичного сигнала $F_с=1{,}5v$, Гц.

1.5. ДЕМОДУЛЯЦИЯ И ДЕКОДИРОВАНИЕ

Восстановление переданного сообщения в приемнике обычно осуществляется в такой последовательности. Сначала производится демодуляция сигнала. В системах передачи непрерывных сообщений в результате демодуляции восстанавливается первичный сигнал, отображающий переданное сообщение. Этот сигнал затем поступает на воспроизводящее или записывающее устройство. В радиовещании таким устройством может быть громкоговоритель или магнитофон. В системах передачи дискретных сообщений в результате демодуляции последовательность элементов сигнала превращается в последовательность кодовых символов, после чего эта последовательность преобразуется в последовательность элементов сообщения, выдаваемую получателю. Это преобразование называется декодированием.

Не следует думать, что демодуляция и декодирование — это просто операции, обратные модуляции и кодированию, выполняемые над пришедшим из канала сигналом. В результате различных искажений и воздействия помех пришедший сигнал может существенно отличаться от переданного. Поэтому всегда можно высказать ряд предположений (гипотез) о том, какое сообщение передавалось. Задачей приемного устройства является принятие решения о том, какое из возможных сообщений действительно передавалось источником. Для принятия такого решения принятый сигнал подвергается анализу с учетом всех сведений об источнике (например, о вероятностях, с которыми источник посылает то или иное сообщение), о применяемом коде и методе модуляции, а также о свойствах канала. В результате такого анализа обычно можно определить условные или апостериорные вероятности всех возможных гипотез и на основании этих вероятностей принять решение, которое и поступает к получателю. Та часть приемного устройства, которая осуществляет анализ приходящего сигнала и принимает решение о переданном сообщении, называется *решающей схемой*.

В системах передачи непрерывных сообщений при аналоговой модуляции решающая схема определяет по пришедшему искаженному вторичному сигналу наиболее вероятный переданный первичный сигнал и восстанавливает его. Здесь решающей схемой является демодулятор. В системах передачи дискретных сообщений решающая схема обычно состоит из двух частей: первой решающей схемы — демодулятора и второй решающей схемы — декодера.

Иногда операции демодуляции и декодирования выполняет одно устройство, которое приходящую последовательность элементов сигнала преобразует сразу в последовательность букв сообщения. Такой метод приема называют *«приемом в целом»*, в отличие от «поэлементного приема» с двумя решающими схемами. В первом случае анализируется целиком отрезок сигнала, соответствующий кодовой комбинации, и на основании того или иного критерия восстанавливается переданный элемент сообщения (буква). Во втором случае сначала анализируются отдельные элементы сигнала, соответствующие кодовым символам, а затем восстановленная кодовая комбинация декодируется, т. е. преобразуется в элемент (букву) сообщения.

В некоторых случаях роль решающей схемы выполняет полностью или частично человек. Так, например, при приеме телеграфных сигналов на слух оператор решает, какой сигнал («точка» или «тире») был передан. Он же выполняет и операцию декодирования.

В приемниках дискретных сообщений, предназначенных для записи информации, все указанные операции выполняются автоматически. В этих случаях приемник принимает решение, какому переданному сигналу соответствует искаженный сигнал.

В простейшем случае первая решающая схема представляет собой пороговое устройство в форме реле, триггера, работающих по принципу «да» или «нет». Если принятый элемент сигнала выше порога, выдается один символ кода (например, 1), если ниже — другой (0).

В некоторых случаях применяются решающие схемы с двумя порогами. При попадании уровня сигнала между двумя порогами решение не принимается — вместо сомнительного элемента сигнала выдается специальный символ стирания. Введение такого стирающего символа облегчает возможность правильного декодирования принятой кодовой комбинации.

Для принятия решения о том, какое сообщение передавалось, необходимо проанализировать пришедший сигнал. С этой целью он подвергается различным преобразованиям, которые называют *обработкой сигнала*. Одной из задач теории связи является отыскание правил оптимальной обработки сигнала, при которой решение о переданном сообщении оказывается наиболее достоверным. Эти правила зависят от свойств канала и методов передачи (кодирования и модуляции). Иногда оптимальные правила обработки оказываются сложными и с целью упрощения аппаратуры используют другую, не оптимальную обработку.

Операции, входящие в процедуру обработки сигнала, могут быть весьма разнообразными. Чаще всего приходится применять в процессе обработки фильтрацию, перемножение сигналов, сложение, интегрирование (реже дифференцирование), стробирование, ограничение, сравнение двух или нескольких отсчетов, возведение в квадрат и другие функциональные преобразования. Многие операции обработки принятого сигнала направлены на то, чтобы увеличить отношение мощности сигнала к мощности помехи.

1.6. ДИСКРЕТИЗАЦИЯ И КОДИРОВАНИЕ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Как уже отмечалось в § 1.1, непрерывные сообщения можно передавать дискретными сигналами. Во многих отношениях такой метод передачи оказывается предпочтительным и в последние годы находит все большее применение. Операция преобразования непрерывного сообщения в дискретное называется *дискретизацией.*

Дискретизация осуществляется не только по времени (как в импульсных методах модуляции), но и по уровням (состояниям). Дискретизация по времени выполняется путем взятия отсчетов функции $b(t)$ в определенные дискретные моменты времени t_k. В результате непрерывная функция $b(t)$ заменяется совокупностью мгновенных значений $\{b_k\} = \{b(t_k)\}$. Обычно моменты отсчетов выбираются на оси времени равномерно, т. е. $t_k = k\Delta t$.

В некоторых случаях сообщение может представлять собой функцию не одного, а нескольких переменных. Примером такого сообщения является телевизионное изображение, которое можно представить как функцию $b(x, y, t)$ двух пространственных координат, x и y, и времени t, где b — яркость точки изображения. Дискретизация по времени осуществляется с помощью кадровой развертки. Шаг дискретизации Δt определяется числом кадров в секунду. В результате строчной развертки дискретизируется координата y, координата x при этом остается непрерывной. Шаг дискретизации Δy определяется числом строк развертки. Таким образом, получается функция

$$b_{ik}(t) = b(i\Delta x, k\Delta t, vt),$$

где v — скорость развертки вдоль строки; i — номер строки; k — номер кадра.

Дискретизация значений функции (уровня) носит название *квантования.* Операция квантования сводится к тому, что вместо данного мгновенного значения передаваемого сообщения (или первичного сигнала) $b(t)$ передаются ближайшие значения по установленной шкале дискретных уровней. Само собой разумеется, что при квантовании вносится погрешность, так как истинные значения b заменяются округленными значениями b_k. Дискретизация по времени лежит в основе всех видов импульсной модуляции. Дискретизация по времени и уровню позволяет непрерывное сообщение преобразовать в дискретное, которое затем кодируется. Достоинством систем связи с дискретизацией являются возможность применения кодирования для повышения помехоустойчивости, удобство обработки сигналов и сопряжения устройств связи с цифровыми вычислительными машинами. Подробно кодовые способы модуляции рассматриваются в гл. 7.

1.7. ОСНОВНЫЕ ХАРАКТЕРИСТИКИ СИСТЕМЫ СВЯЗИ

При оценке работы системы связи необходимо, прежде всего, учесть, какую точность передачи сообщения обеспечивает система и с какой скоростью передается информация. Первое определяет качество передачи, второе — количество.

В реальной системе связи качество передачи зависит от степени искажений принятого сообщения. Эти искажения зависят от свойств и технического состояния системы, а также от интенсивности и характера помех. В правильно спроектированной и технически исправной системе связи искажения сообщений обусловлены лишь воздействием помех. В этом случае качество передачи полностью определяется помехоустойчивостью системы.

Под *помехоустойчивостью* обычно понимают способность системы противостоять вредному влиянию помех на передачу сообщений. Так как действие помех проявляется в том, что принятое сообщение отличается от переданного, то количественно помехоустойчивость при заданной помехе можно характеризовать степенью соответствия принятого сообщения переданному. Назовем эту величину общим термином — *верность*. Количественную меру верности приходится выбирать по-разному в зависимости от характера сообщения.

Пусть сообщение представляет собой дискретную последовательность элементов из некоторого конечного множества. Влияние помехи на передачу такого сообщения проявляется в том, что вместо фактически переданного элемента может быть принят какой-либо другой, такое событие называется ошибкой. В качестве количественной меры верности можно взять вероятность ошибки p или любую монотонную функцию этой вероятности.

При передаче непрерывных сообщений степенью соответствия принятого сообщения $\hat{b}(t)$ переданному $b(t)$ может служить некоторая величина ε, представляющая собой «расстояние» между $\hat{b}$ и b. Часто принимается критерий квадратичного отклонения, выражающийся соотношением

$$\varepsilon^2 = \frac{1}{T}\int_0^T [\hat{b}(t) - b(t)]^2\, dt. \tag{1.13}$$

Количественную меру верности можно также определить как вероятность того, что уклонение ε не превзойдет некоторой заранее заданной величины ε_0:

$$Q = P(|\varepsilon| \leqslant \varepsilon_0). \tag{1.14}$$

Как будет показано в последующих главах, верность передачи зависит от отношения сигнала к помехе[1]. Чем больше это отношение, тем меньше (при прочих равных условиях) вероятность ошибки (больше вериость).

При данной интенсивности помехи вероятность ошибки тем меньше, чем сильнее различаются между собой сигналы, соответствующие разным сообщениям. Задача состоит в том, чтобы выбрать для передачи сигналы с бо́льшим различием. Наконец, верность передачи зависит и от способа приема. Нужно выбрать такой способ приема, который наилучшим образом реализует

[1] Точное определение этого отношения будет дано в гл. 4 и 6.

различие между сигналами при данном отношении сигнала к помехе.

Необходимо обратить внимание на существенное различие между аналоговыми и дискретными системами передачи сообщений. В аналоговых системах всякое, даже сколь угодно малое мешающее воздействие на сигнал, вызывающее искажение модулируемого параметра, всегда влечет за собой внесение соответствующей ошибки в сообщение. Поэтому абсолютно точное восстановление переданного сообщения невозможно. В дискретных системах ошибка при передаче сообщений возникает только тогда, когда сигнал опознается неправильно, а это происходит лишь при сравнительно больших искажениях.

В теории помехоустойчивости, разработанной В. А. Котельниковым, показывается, что при заданном методе модуляции существует предельная (потенциальная) помехоустойчивость, которая ни при каком способе приема не может быть превзойдена. Приемное устройство, реализующее потенциальную помехоустойчивость, называется оптимальным приемником.

Наряду с верностью важнейшим показателем работы системы связи является *скорость передачи.* В системах передачи дискретных сообщений скорость измеряется числом передаваемых двоичных символов в секунду R. Для одного канала скорость передачи определяется соотношением (см. ниже гл. 2)

$$R = \frac{1}{T} \log_2 m, \tag{1.15}$$

где T — длительность элементарной посылки сигнала; m — основание кода. При $m=2$ имеем $R=1/T=v$, Бод.

Максимальную скорость передачи $R_{\text{макс}}$, допускаемую данной системой связи при условии, что канал не вносит ошибок и искажений, принято называть *пропускной способностью системы.* Пропускную способность системы передачи аналоговых сообщений оценивают количеством одновременно передаваемых телефонных разговоров, радиовещательных или телевизионных программ и т. п.

Пропускную способность системы $R_{\text{макс}}$ не следует путать с *пропускной способностью канала связи* C. Точное определение пропускной способности канала будет дано в гл. 3. Пока лишь отметим, что она характеризует максимальное количество информации, которое может быть передано по данному каналу в единицу времени.

В реальных системах скорость передачи R всегда меньше пропускной способности канала C. В теории информации доказывается, что при $R<C$ существуют такие способы передачи и соответствующие способы приема, при которых верность передачи может быть сделана сколь угодно большой. Соответствующие теоремы будут изложены в гл. 3 и 6.

Пропускная способность системы связи — понятие техническое, характеризующее используемую аппаратуру, тогда как про-

пускная способность канала является фундаментальным теоретическим понятием, определяющим потенциальные возможности системы связи, использующей данный канал, если на сложность и стоимость аппаратуры не наложено никаких ограничений и к тому же допускается любая задержка переданных сообщений.

Под *задержкой* понимается максимальное время, прошедшее между моментом подачи сообщения от источника на вход передающего устройства и моментом выдачи восстановленного сообщения приемным устройством. Задержка является также одной из важных характеристик системы связи. Она зависит, во-первых, от характера и протяженности канала, во-вторых, от длительности обработки сигнала в передающем и приемном устройствах. Последняя определяется, главным образом, количеством информации, объединяемой в одну кодовую последовательность. Действительно, пока источник не выдаст всей этой информации, процесс кодирования не может начаться, а пока вся кодовая комбинация не будет принята, не может начаться процесс декодирования. Скорость передачи и задержка являются независимыми характеристиками, практически не связанными друг с другом.

Существуют и многие другие параметры, характеризующие с различных точек зрения качество системы связи. К ним, в частности, относятся скрытность связи, надежность системы, габариты и масса аппаратуры, стоимость оборудования, эксплуатационные расходы и т. п. Эти характеристики в курсе «Теория передачи сигналов» не рассматриваются. Им посвящены отдельные разделы других специальных курсов.

ГЛАВА ВТОРАЯ

СООБЩЕНИЯ, СИГНАЛЫ, ПОМЕХИ

2.1. СООБЩЕНИЯ, СИГНАЛЫ И ПОМЕХИ КАК СЛУЧАЙНЫЕ ПРОЦЕССЫ

Как уже отмечалось в § 1.1, детерминированное, т. е. заранее известное, сообщение не содержит информации. Поэтому в теории связи источник сообщения следует рассматривать как устройство, осуществляющее выбор из некоторого множества возможных сообщений. Каждая конкретная реализация сообщения возникает с определенной вероятностью, которая в общем случае зависит от того, какие сообщения передавались раньше. Точно так же и посылаемая в канал реализация сигнала является элементом некоторого множества, выбираемого с определенной вероятностью.

Множество, на котором задана вероятностная мера, называют *ансамблем*. Ансамбли сообщений и сигналов могут быть конечными (в дискретном случае) или бесконечными.

Ансамбль $\{X(t)\}$ функций времени является *случайным процессом*. Каждая входящая функция $x_r(t)$ называется *выборочной функцией* или *реализацией* процесса. Наличие различных реализаций позволяет сигналу переносить информацию. Для этого нужно установить соответствие между каждым сообщением из ансамбля сообщений и определенной реализацией сигнала. Тогда по принятой реализации сигнала можно судить о том, какое сообщение передавалось. Это суждение было бы точным, если бы переданный сигнал не искажался помехами. В действительности помехи являются также случайными процессами, и поэтому по принятому сигналу можно лишь с некоторой вероятностью определить, какое сообщение было передано.

Итак, сообщения, сигналы, помехи являются случайными процессами и для их исследования необходимо использовать основные положения теории случайных процессов (случайных функций). Все содержание данной главы относится в равной степени к случайным сообщениям, сигналам и помехам [1].

Случайный процесс может быть задан на дискретном множестве значений t: $t_1, t_2, \ldots$ Это случай сообщения (сигнала), дискретного по времени. Такие случайные процессы называются также *случайными последовательностями*. Примером случайной последовательности является процесс $X(t_k)$, заданный на дискретных точках $t_1, t_2, \ldots, t_k, \ldots$ и принимающий в каждой из них значение 1 с вероятностью p_1 и 0 с вероятностью $p_0 = 1 - p_1$, независимо от значений в других точках. Чаще встречаются процессы, непрерывные по времени, например, заданные на всей оси $-\infty < t < \infty$ или на конечном отрезке $0 < t < \pi$, $-T/2 < t < T/2$ и т. п. Процессы, заданные на конечном отрезке времени, называются *финитными*.

Задать детерминированную (т. е. не случайную) функцию времени — это значит указать ее значение для любого t в пределах области определения. Случайную функцию можно задать только в вероятностном смысле. Если число реализаций конечно или счетно, можно их попросту перечислить и указать их вероятности.

Рассмотрим, как задаются дискретные случайные последовательности элементов $A(t_k)$, могущие принимать m различных значений $a_1, a_2, \ldots, a_m$. Реализацией такой последовательности может быть, например, $A(t_1) = a_4^{(1)}$; $A(t_2) = a_6^{(2)}$; $A(t_3) = a_1^{(3)}$; ..., $A(t_k) = a_{i_k}^{(k)}$; ..., где нижний индекс обозначает значение элемента, а верхний — момент времени. Отрезок такой последовательности длины n (т. е. состоящий из n элементов) характеризуется своей вероятностью P. На основании теоремы умножения вероятностей можно в общем случае записать вероятность

---

[1] Как уже отмечалось, сообщение может быть функцией не только времени, но и других переменных (например, пространственных координат). Тем не менее, будем обозначать аргумент случайного процесса, как правило, буквой t.

отрезка реализации, в которой элементы $A(t_k)$ принимают значения a_{i_k} ($i_k=1, 2, \ldots, m$) так:

$$P\left[a_{i_1}^{(1)}, a_{i_2}^{(2)}, a_{i_3}^{(3)}, \ldots, a_{i_n}^{(n)}\right)=P\left(a_{i_1}^{(1)}\right)P\left(a_{i_2}^{(2)}\,|\,a_{i_1}^{(1)}\right)P\left(a_{i_3}^{(3)}\,|\,a_{i_1}^{(1)}, a_{i_2}^{(2)}\ldots\right)$$

$$P\left(a_{i_n}^{(n)}\,|\,a_{i_1}^{(1)}, a_{i_2}^{(2)}, \ldots, a_{i_{n-1}}^{(n-1)}\right). \qquad (2.1)$$

Здесь $P(a_{i_n}^{(n)}|a_{i_1}^{(1)}, a_{i_2}^{(2)}, \ldots, a_{i_{n-1}}^{(n-1)})$ условная вероятность появления элемента a_{i_n} в момент t_n при условии, что в предыдущие моменты осуществлялась реализация отрезка $a_{i_1}, a_{i_2}, \ldots, a_{i_{n-1}}$.

Примером такой последовательности может служить последовательность букв, передаваемых по телеграфу. Хотя для этого случая определить вероятность отрезка нелегко, но качественно сравнить вероятности различных отрезков можно без особого труда. Так, при $n=7$ последовательность букв «ВЫЕЗЖАЮ» более вероятна, чем последовательность «ГРАНАТА», которая, в свою очередь, вероятнее, чем последовательность «МАФТУЗА». Последняя все же вероятнее, чем последовательность «АЬМШЗЮО», для которой вероятность появления в осмысленном русском тексте равна нулю. На этом примере, между прочим, легко убедиться, что условная вероятность появления некоторой буквы зависит от ранее переданных букв.

Простейший вид случайной последовательности элементов отличается тем, что появление того или иного элемента не зависит от предыдущих элементов. Для такой последовательности независимых элементов (или последовательности Бернулли)

$$P\left(a_{i_1}^{(1)}, a_{i_2}^{(2)}, \ldots, a_{i_n}^{(n)}\right)=P\left(a_{i_1}\right)P\left(a_{i_2}\right)\cdots P\left(a_{i_n}\right). \qquad (2.2)$$

Важным видом случайной последовательности зависимых элементов является *цепь Маркова*. Простой цепью Маркова называется последовательность, в которой условная вероятность появления некоторого элемента a_{ik} целиком определена, если известен предыдущий элемент $a_{i,\,k-1}$. Это значит, что

$$P\left(a_{i_k}^{(k)}\,|\,a_{i1}^{(1)}, a_{i_2}^{(2)}, \ldots, a_{i_{k-1}}^{(k-1)}\right)=P\left(a_{i_k}^{(k)}\,|\,a_{i_{k-1}}^{(k-1)}\right). \qquad (2.3)$$

Таким образом, в простой цепи Маркова связь между последовательными элементами целиком определяется зависимостью между соседними элементами. Для вычисления вероятности некоторого отрезка цепи Маркова достаточно знать переходные вероятности, т. е. все условные вероятности $P(a_i|a_j)$ появления элемента a_i, если предыдущим элементом был a_j. Заметим, что это вовсе не означает независимости $a_{i_k}^{(k)}$ от $a_{i_{k-2}}^{(k-2)}$ или от еще более ранних элементов. Поскольку вероятность появления $a_{i_k}^{(k)}$ зависит от того, каким был элемент $a_{i_{k-1}}^{(k-1)}$, а последний зависит от $a_{i_{k-2}}^{(k-2)}$, то имеет место и зависимость $a_{i_k}^{(k)}$ от $a_{i_{k-2}}^{(k-2)}$, если значение $a_{i_{k-1}}^{(k-1)}$ неизвестно. Нетрудно показать, что

$$P\left(a_{i_k}^{(k)}\,|\,a_{i_{k-2}}^{(k-2)}\right)=\sum_{i_{k-1}=1}^{m}P\left(a_{i_k}^{(k)}\,|\,a_{i_{k-1}}^{(k-1)}\right)P\left(a_{i_{k-1}}^{(k-1)}\,|\,a_{i_{k-2}}^{(k-2)}\right). \qquad (2.4)$$

Обобщением простой цепи Маркова является сложная марковская цепь r-го порядка, в которой вероятность элемента целиком определена, если известны r предшествующих ему элементов. Марковская цепь высокого порядка является удовлетворительной математической моделью текстовых последовательностей в приведенном выше примере.

Непрерывные случайные функции также могут иметь конечное число реализаций на конечном интервале времени. Для их задания также достаточно указать их вероятности.

Так, например, процесс $X(t)$, заданный на интервале $0 \leqslant t \leqslant T$ и имеющий две реализации:

$$x_1(t) = A\cos\omega_1 t; \quad x_2(t) = A\cos\omega_2 t\}, \tag{2.5}$$

возникающие с вероятностями $P(x_1)=p_1$, $P(x_2)=p_2=1-p_1$, где A, ω_1, ω_2, p_1 — определенные постоянные, является простым примером случайного процесса с конечным числом реализаций. Он может служить сигналом, информационное содержание которого определяется значением частоты ω.

Примером процесса, заданного на том же интервале и имеющего бесконечное число реализаций, может служить процесс

$$X(t) = A\cos(\omega t + \Phi), \tag{2.6}$$

где A и ω — постоянные неслучайные величины, а Φ — случайная величина с заданным распределением вероятностей. Такой процесс может иметь счетное число реализаций (например, если Φ принимает значения $\varphi_k = 2\pi/k$ с вероятностями 2^{-k}) или несчетное (например, если величина Φ равномерно распределена на интервале 0—2π). Довольно часто приходится рассматривать процессы, определенные рядом

$$X(t) = \sum_{k=1}^{\infty} X_k \psi_k(t), \tag{2.7}$$

где $\psi_k(t)$ — обычные (детерминированные) функции, а X_k — случайные величины с заданным распределением вероятности. Сходимость ряда понимается в среднеквадратичном смысле, т. е. при всех t:

$$\lim_{n\to\infty} M\left\{\left|X(t) - \sum_{k=1}^{n} X_k \psi_k(t)\right|^2\right\} = 0. \tag{2.8}$$

Символ M{ } здесь и в дальнейшем обозначает математическое ожидание.

В общем случае скалярный случайный процесс $X(t)$ полностью задан, если для любого набора n моментов времени $t_1, t_2, \ldots, t_n$ и любых значений $x_1, x_2, \ldots, x_n$ можно вычислить вероятность того, что $X(t)$ принимает в указанные моменты времени значения, не превышающие соответственно $x_1, x_2, \ldots, x_n$:

$$F(x_1, x_2, \cdots, x_n; t_1, t_2, \cdots, t_n) = P\{X(t_1) \leqslant x_1, X(t_2) \leqslant x_2, \cdots, X(t_n) \leqslant x_n\}. \tag{2.9}$$

Здесь $P\{\ \}$ обозначает совместную вероятность событий, записанных в скобках. Заметим, что $X(t_k)$ представляет собой случайную величину и называется *сечением* случайного процесса в момент t_k.

Функция $F(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n)$ в (2.9) называется n-мерной *функцией распределения вероятности* процесса. Ее аргументами являются $x_1, x_2, \ldots, x_n$. Величины $t_1, t_2, \ldots, t_n$ являются пара-

метрами. Следует подчеркнуть, что случайный процесс полностью задан в том случае, если для любого n и для любых моментов $t_1, t_2, \ldots, t_n$ (в области его определения) можно найти функцию распределения.

На рис. 2.1 представлены четыре реализации некоторого случайного процесса и показаны три его сечения. Реализации *1* и *2* удовлетворяют условию $X(t_1) \leqslant x_1$, $X(t_2) \leqslant x_2$, $X(t_3) \leqslant x_3$, где x_1, x_2, x_3 — значения сечений, отмеченные на рисунке точками, а реализации *3* и *4* этому условию не удовлетворяют. Если других реализаций данный процесс не имеет, то трехмерная функция распределения при заданных сечениях x_1, x_2, x_3 равна сумме вероятностей реализаций *1* и *2*.

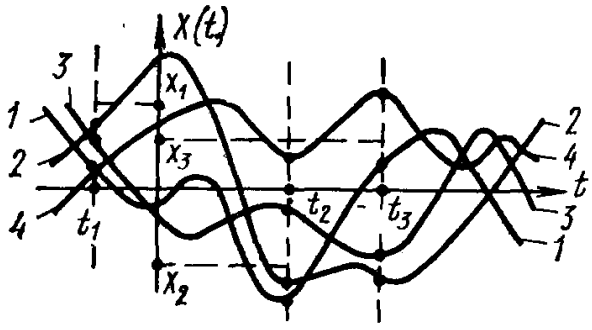


Рис. 2.1. Реализация случайного процесса

Если существуют частные производные функции распределения по всем x_k, то можно определить n-мерную *плотность распределения вероятности:*

$$w(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = \frac{\partial^n F(x_1, \ldots, x_n; t_1, \ldots, t_n)}{\partial x_1 \partial x_2 \ldots \partial x_n}, \tag{2.10}$$

которая также полностью определяет процесс, если она известна для любого числа любых сечений [1].

Пусть например,

$$w(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = A_n \exp\left\{ -\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} [(x_i - a_i) \times \right.$$

$$\left. \times (x_j - a_j)] \right\}, \tag{2.11}$$

где A_n, c_{ij}, a_i, a_j — постоянные, определенные выбором сечений $t_1, \ldots, t_n$ и связанные между собой некоторыми зависимостями, на которых сейчас останавливаться не будем. В частности, при

---

[1] С помощью обобщенной функции $\delta(x)$ можно распространить понятие плотности распределения и на случай разрывной функции распределения. Как известно, если функция $F(x)$ при некотором значении $x = x_0$ совершает скачок на величину a, то в ее производную войдет слагаемое $a\delta(x - x_0)$. Функция $\delta(x)$ равна бесконечности при $x = 0$ и равна нулю при прочих значениях x, причем при $x_2 > x_1$

$$\int_{x_1}^{x_2} \delta(x - x_0)\, dx = \begin{cases} 0, & \text{если } x_0 < x_1 \text{ или } x_0 > x_2; \\ 1, & \text{если } x_1 < x_0 < x_2. \end{cases} \tag{2.12}$$

$n=1$ одномерную плотность распределения можно записать следующим образом:

$$w(x,\ t)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right). \tag{2.13}$$

Здесь $1/\sqrt{2\pi}\sigma$ соответствует коэффициенту A_1, а $1/2\sigma^2$ — коэффициенту c_{11} в формуле (2.11).

Процесс, у которого любая n-мерная плотность распределения вероятностей выражается формулой (2.11), называется *нормальным*, или *гауссовским*. Нормальные процессы играют исключительно важную роль в теории передачи сигналов. Так, в § 1.3 уже говорилось, что флуктуационные помехи, в частности, вызванные тепловым движением электронов, являются нормальными. Из центральной предельной теоремы вероятности известно, что при некоторых ограничениях сумма большого числа разных процессов образует процесс, очень близкий к гауссовскому.

Часто случайный процесс можно определить через другой случайный процесс с известными функциями распределения, например

$$X(t)=Y_1(t)\,s(t)+Y_2(t)\,\tilde{s}(t), \tag{2.14}$$

где Y_1 и Y_2 — нормальные процессы, а $s(t)$ и $\tilde{s}(t)$ — заданные регулярные функции. Другой пример:

$$X(t)=\int_{-\infty}^{t}Y(\theta)\,g(t-\theta)\,d\theta, \tag{2.15}$$

где $Y(t)$ — случайный процесс с известными функциями распределения, а $g(t)$ детерминированная функция [1]. Иногда удобно задать случайный процесс $X(t)$ дифференциальным уравнением, например [2]

$$\frac{dX(t)}{dt}+\alpha(t)\,X(t)=N(t), \tag{2.16}$$

где $N(t)$ — известный случайный процесс, а $\alpha(t)$ — детерминированная функция.

Среднее значение процесса по ансамблю или его математическое ожидание определяется как

$$\overline{X(t)}=M\{X(t)\}=\int_{-\infty}^{\infty}x\,w(x,\ t)\,dx, \tag{2.17}$$

где $w(x,\ t)$ — одномерная плотность распределения для сечения t. Математическое ожидание, вообще говоря, является функцией времени (не случайной). Разность между случайным процессом и

---

[1] Здесь и далее интегрирование случайного процесса понимается в среднеквадратичном смысле [см. формулу (2.8)].

[2] Производной $Y(t)=dX/dt$ случайного процесса $X(t)$ (в среднеквадратичном смысле) называется процесс, удовлетворяющий условию

$$\lim_{\Delta t\to 0}M\left\{\left|\frac{X(t+\Delta t)-X(t)}{\Delta t}-Y(t)\right|^2\right\}=0.$$

его математическим ожиданием называется центрированным процессом и обозначается $\overset{\circ}{X}(t)=X(t)-\overline{X(t)}$. Математическое ожидание квадрата центрированного процесса называется *дисперсией*:

$$\overline{[\overset{\circ}{X}(t)]^2}=\mathrm{D}\{X(t)\}=\int_{-\infty}^{\infty}[x-\overline{X(t)}]^2 w(x,\ t)\,dx, \tag{2.18}$$

которая, вообще говоря, тоже является функцией времени.

Функция корреляции $\mathrm{B}_X(t_1,t_2)$ определяется как математическое ожидание произведения двух сечений центрированного случайного процесса:

$$\overline{\overset{\circ}{X}(t_1)\overset{\circ}{X}(t_2)}=\mathrm{B}_X(t,\ t_2)=$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x_1-\overline{X(t_1)}]\,[x_2-\overline{X(t_2)}]\,w(x_1,\ x_2;\ t_1,\ t_2)\,dx_1dx_2. \tag{2.19}$$

Функция корреляции (ФК), вообще говоря, является функцией двух моментов времени, t_1 и t_2. Ее называют также функцией автокорреляции, в отличие от функции взаимной корреляции между двумя процессами $X(t)$ и $Y(t)$:

$$\overline{\overset{\circ}{X}(t_1)\overset{\circ}{Y}(t_2)}=\mathrm{B}_{XY}(t_1,\ t_2)=$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x_1-\overline{X(t_1)}]\,[y_2-\overline{Y(t_2)}]\,w(x_1,\ y_2;\ t_1,\ t_2)\,dx_1dx_2,$$

где $w(x_1,\ y_2;\ t_1,\ t_2)$ — двумерная (совместная) плотность вероятности сечения t_1 процесса X и сечения t_2 процесса Y. Иногда функции автокорреляции и взаимной корреляции определяют без центрирования.

Легко показать, что для нормального процесса с одномерной плотностью (2.12) математическое ожидание равно a, а дисперсия — σ^2.

В теории вероятности доказывается, что все коэффициенты n-мерной функции распределения нормального процесса (2.11) выражаются через математическое ожидание, дисперсию и функцию корреляции, которые, в свою очередь, можно определить, зная двумерное распределение. Отсюда вытекает замечательное свойство, присущее только нормальному процессу, он полностью характеризуется своим двумерным распределением.

Случайный процесс, у которого математическое ожидание и дисперсия не зависят от времени, а ФК зависит от разности $t_2-t_1=\tau$, но не от самих значений t_1 и t_2, называется *стационарным* (в широком смысле) [1]. Реальные сообщения, сигналы и по-

[1] Стационарными в узком смысле называются процессы, у которых любые n-мерные функции распределения (2.2) не изменяются, если все моменты времени t_1, ..., t_n сдвинуть на любой интервал τ. Очевидно, что процесс, стационарный в узком смысле, является стационарным и в широком смысле. Обратное утверждение неверно. В настоящей книге стационарность понимается всегда в широком смысле.

мехи обычно не являются стационарными. Однако если их рассматривать на протяжении не слишком длительного времени, то с хорошим приближением их можно описать стационарными процессами, которые поэтому широко используются в качестве математической модели реальных сообщений, сигналов и помех.

Примером стационарного процесса может служить процесс (2.6), если Φ — случайная величина, распределенная равномерно на интервале $(0; 2\pi)$. Его математическое ожидание равно нулю, дисперсия равна $0{,}5A^2$, а функция корреляции $B_X(t_1, t_2) = \frac{1}{2}A^2 \cos[\omega(t_2 - t_1)] = 0{,}5A^2 \cos \omega\tau$. Наоборот, процесс (2.5) стационарным не является. Заметим, что процесс (2.15), вообще говоря, не является стационарным, даже если стационарен процесс $Y(t)$.

Помимо средних значений по ансамблю можно определить средние значения случайного процесса по времени. Для финитного процесса, заданного на интервале времени от t_1 до t_2, определим *постоянную составляющую*:

$$\overline{\overline{X(t)}} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} X(t)\, dt. \tag{2.20}$$

Здесь и в дальнейшем волнистая черта означает усреднение по времени, в отличие от прямой черты, которой обозначается статистическое усреднение.

Если процесс задан на бесконечной оси времени, то для него постоянная составляющая определяется как

$$\overline{\overline{X(t)}} = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} X(t)\, dt. \tag{2.20а}$$

Постоянная составляющая, очевидно, сама от времени не зависит, но является, вообще говоря, случайной величиной — ее значение зависит от реализации сигнала.

Процесс $X(t) - \overline{\overline{X(t)}} = X_{\sim}(t)$ называется переменной составляющей процесса $X(t)$. Среднее по времени значение квадрата переменной составляющей

$$\overline{\overline{X_{\sim}^2(t)}} = \begin{cases} \dfrac{1}{t_2 - t_1} \displaystyle\int_{t_1}^{t_2} X_{\sim}^2(t)\, dt — \text{для финитного сигнала;} \\ \lim\limits_{T \to \infty} \dfrac{1}{T} \displaystyle\int_{-T/2}^{T/2} X_{\sim}^2(t)\, dt — \text{для нефинитного сигнала} \end{cases} \tag{2.21}$$

также является случайной величиной, не зависящей от времени. Ее обычно называют *мощностью* переменной составляющей, что совпадает с обычным определением мощности в электротехнике, если процесс представляет собой ток, проходящий через единичное сопротивление, или напряжение, приложенное к единичному сопротивлению.

Стационарные процессы называются *эргодическими*, если для них усреднение по времени приводит к тем же результатам, что и статистическое усреднение, т. е. математическое ожидание равно постоянной составляющей, а дисперсия — мощности переменной составляющей. Грубо говоря, эргодичность процесса заключается в том, что все его реализации похожи друг на друга. Эргодические процессы часто используются в качестве математических моделей реальных сообщений, сигналов и помех.

Функцию корреляции эргодического процесса также можно вычислить по одной реализации усреднением по времени:

$$B_X(\tau) = \overline{x(t)\,x(t+\tau)} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)\,x(t+\tau)\,dt. \qquad (2.22)$$

Рассмотрим некоторые свойства функции корреляции стационарного случайного процесса, которую будем обозначать $B_X(\tau)$, где τ — разность между двумя сечениями. С учетом стационарности можно записать:

$$B_X(\tau) = \overline{\mathring{X}(t)\mathring{X}(t+\tau)} = \overline{\mathring{X}(t-\tau)\mathring{X}(t)} = B_X(-\tau). \qquad (2.23)$$

Таким образом, ФК стационарного процесса является четной функцией интервала τ. Положив $\tau=0$, найдем:

$$B_X(0) = \overline{\mathring{X}^2(t)} = D\{X(t)\}, \qquad (2.24)$$

т. е. значение корреляционной функции в точке $\tau=0$ равно дисперсии сигнала.

В теории вероятностей доказывается, что при любом τ $|B_X(\tau)| \leqslant D\{X(t)\}$.

Таким образом, $B_X(\tau)$ имеет максимум при $\tau=0$.

Определим *нормированную функцию корреляции* (НФК) или коэффициент корреляции процесса $X(t)$ как

$$R_X(t_1, t_2) = B_X(t_1\ t_2)/\sqrt{D\{X(t_1)\}\,D\{X((t_2)\}}. \qquad (2.25)$$

В частности, для стационарного процесса

$$R_X(\tau) = B_X(\tau)/D\{X\}. \qquad (2.26)$$

Из свойств функции корреляции следует, что для стационарного процесса

$$R_X(\tau) = R_X(-\tau),\ R_X(0) = 1,\ |R_X(\tau)| \leqslant 1. \qquad (2.27)$$

Величина $R_X(\tau)$ является в известной степени мерой статистической зависимости между сечениями процесса, отстоящими на интервал τ. Действительно, если при некотором значении τ $X(t)$ и $X(t+\tau)$ всегда совпадают, то, как легко убедиться, $R(\tau)=1$. Если при другом значении τ $X(t)=-X(t+\tau)$, то $R(\tau)=-1$. Если, наконец, $X(t)$ и $X(t+\tau)$ являются независимыми случайными величинами, то $B_X(\tau)=M\{\mathring{X}(t)\mathring{X}(t+\tau)\}=M\{\mathring{X}(t)\}\times M\{\mathring{X}(t+\tau)\}=0$, откуда и $R_X(\tau)=0$.

Таким образом, независимые сечения являются также некоррелированными. Обратное утверждение в общем случае неверно — два сечения могут быть некоррелированными и в то же время зависимыми. Только в том случае, когда процесс гауссовский, из некоррелированности двух сечений следует их независимость.

Коэффициент и функция корреляции для эргодических случайных процессов стремятся к нулю с увеличением τ. Это происходит потому, что чем дальше отстоят друг от друга сечения, тем слабее статистическая зависимость между ними. Интервал времени $\tau_к$ называется временем корреляции процесса, если при $\tau > \tau_к$ величина $|R(\tau)|$ становится пренебрежимо малой (практически в большинстве случаев для этого достаточно, чтобы она стала меньше 0,1). Иногда интервал корреляции определяют как основание прямоугольника с высотой, равной 1, площадь которого равна площади, ограниченной кривой $|R(\tau)|$ и осями координат:

$$\tau_к = \int_0^\infty |R(\tau)|\, d\tau,$$

если этот интеграл существует.

Многомерные функции и плотности распределения (2.9), (2.10) иногда удобнее выражать через условные функции (плотности) распределения, аналогично тому, как это было сделано для вероятностей отрезка последовательности (2.1). Особенно удобно такое представление для марковских процессов, которые являются аналогом простых марковских цепей и обладают тем свойством, что при известном значении $X(t_{k-1}) = x_{k-1}$ вероятность значения $X(t_k)$ $(t_k > t_{k-1})$ не зависит от значений в любые более ранние моменты времени:

$$w(x_k;\, t_k \mid x_1,\, x_2,\, \ldots,\, x_{k-1};\, t_1,\, t_2,\, \ldots,\, t_{k-1}) =$$
$$= w(x_k;\, t_k \mid x_{k-1};\, t_{k-1}). \qquad (2.28)$$

Отсюда следует, что для марковского процесса при $t_1 < t_2 < \ldots < t_n$

$$w(x_1,\, \ldots,\, x_n;\, t_1,\, \ldots,\, t_n) = w(x_1;\, t_1)\, w(x_2;\, t_2 \mid x_1;\, t_1) \times$$
$$\times w(x_3;\, t_3 \mid x_2;\, t_2) \cdots w(x_n;\, t_n \mid x_{n-1};\, t_{n-1}). \qquad (2.29)$$

Теория марковских процессов хорошо разработана и широко используется в современной теории связи. В частности, при некоторых дополнительных условиях переходная плотность вероятности $w_1(x_2;\, t_2 | x_1;\, t_1)$, где $t_2 > t_1$, удовлетворяет дифференциальному уравнению 2-го порядка в частных производных:

$$\frac{\partial w(x_2;\, t_2 \mid x_1;\, t_1)}{\partial t_2} = -\frac{\partial}{\partial x_2}[A_1(x_2,\, t_2)\, w(x_2;\, t_2 \mid x_1;\, t_1)] +$$
$$+ \frac{1}{2}\,\frac{\partial^2}{\partial x_2^2}[A_2(x_2,\, t_2)\, w(x_2;\, t_2 \mid x_1;\, t_1)] \qquad (2.30)$$

при начальном условии $w_1(x_2;\ t_1|x_1;\ t_1)=\delta(x_2-x_1)$. Здесь $A_1(x_2,\ t_2)$ называют коэффициентом сноса, а $A_2(x_2,\ t_2)$ — коэффициентом диффузии.

Уравнение (2.30) называют уравнением *Колмогорова—Фоккера—Планка*. Марковские процессы, удовлетворяющие этому уравнению, называют диффузионными. В общем случае это нестационарные процессы. Для стационарного процесса коэффициенты A_1 и A_2 не зависят от t_2.

В зависимости от вида A_1 и A_2 диффузионный марковский процесс может иметь различные распределения вероятностей. В частности, он может быть гауссовским. Имеет место следующая теорема: для того чтобы стационарный гауссовский процесс был марковским, необходимо и достаточно, чтобы его коэффициент корреляции был экспоненциальным:

$$R(\tau)=e^{-\alpha|\tau|}, \tag{2.31}$$

где α — постоянная.

2.2. СПЕКТРЫ СЛУЧАЙНЫХ ПРОЦЕССОВ

При изучении различных преобразований детерминированных непериодических сигналов, в частности, прохождения их через линейные цепи широко используется спектральное представление с помощью преобразования Фурье. Напомним, что преобразованием Фурье от $x(t)$ является комплексная функция от частоты f, обозначаемая обычно $S(i\omega)$ или $S(i2\pi f)$ и равная

$$\int_{-\infty}^{\infty} x(t)\,e^{-i2\pi ft}\,dt, \tag{2.32}$$

если этот интеграл существует.

При изучении случайных сигналов и помех желательно определить спектр, характеризующий не отдельную выборочную функцию, а весь ансамбль в целом. При попытке применить преобразование Фурье (2.32) к случайным процессам возникают различные препятствия, а именно:

1. Как известно, для существования интеграла (2.32) необходимо, чтобы функция $x(t)$ была абсолютно интегрируема или хотя бы интегрируема в квадрате[1]. Последнее означает, что интеграл

$$E_x=\int_{-\infty}^{\infty}|x(t)|^2\,dt \tag{2.33}$$

должен сходиться. Заметим, что интеграл (2.33) представляет собой энергию сигнала (при обычном предположении, что через единичное сопротивление проходит ток $x(t)$ или к нему приложено напряжение $x(t)$. Таким образом, преобразование Фурье сущест-

[1] В этом случае преобразование Фурье существует в смысле среднеквадратичной сходимости.

вует для сигналов с конечной энергией. Но в наиболее интересном случае стационарных эргодических процессов каждая выборочная функция с вероятностью 1 имеет бесконечную энергию. Это следует из того, что такой процесс имеет положительную среднюю мощность, равную его дисперсии, а задан он на бесконечной оси времени. Следовательно, преобразование (2.32) для него не существует.

2. Для тех процессов, которые имеют конечную энергию (а это нестационарные процессы)[1], можно определить преобразование Фурье как случайную комплексную функцию частоты:

$$S_X(\mathrm{i}\,2\pi f) = \int_{-\infty}^{\infty} X(t)\,\mathrm{e}^{-\mathrm{i}\,2\pi f}\,dt. \tag{2.34}$$

Однако она оказывается почти бесполезной для описания ансамбля. Это вызвано тем, что каждая ее реализация является индивидуальной характеристикой одной конкретной реализации и из нее трудно извлечь данные, характеризующие все множество.

Выход из этих затруднений заключается в отбрасывании некоторых параметров спектра, а именно спектра фаз, и построении функции, характеризующей распределение энергии процесса по оси частот.

Построение такой функции проведем сначала для процессов с конечной энергией, для которых существует случайная спектральная плотность $S_X(\mathrm{i}\,2\pi f)$ (2.34). Каждая ее реализация соответствует реализации $x(t)$, и для них, согласно теореме Парсеваля,

$$E_x = \int_{-\infty}^{\infty} |x(t)|^2\,dt = \int_{-\infty}^{\infty} |S_x(\mathrm{i}\,2\pi f)|^2\,df. \tag{2.35}$$

Функция $|S_X(\mathrm{i}\,2\pi f)|^2$ характеризует, таким образом, распределение энергии реализации по оси частот. Назовем ее спектральной плотностью энергии реализации.

Усреднив эту функцию по всем реализациям, получим *спектральную плотность энергии процесса*:

$$M\{|S_x(\mathrm{i}\,2\pi f)|^2\}. \tag{2.36}$$

Отметим некоторые свойства спектральной плотности энергии. Из ее определения следует, что она неотрицательна. Кривая, изображающая эту функцию, охватывает площадь, равную математическому ожиданию энергии процесса. Она определена на бесконечной оси, т. е. как для положительных, так и для отрицательных частот. Как известно, из теории преобразования Фурье, для действительной функции $x(t)$ модуль $|S_x(\mathrm{i}\,2\pi f)|$ является четной функцией частоты. Поэтому о спектральном распределении энергии процесса можно судить и по одной половине графика функции (2.36), например при $f \geqslant 0$.

[1] Множество функций с конечной энергией называют «пространством L_2». Об этом будет подробнее сказано в § 2.5.

Перейдем теперь к стационарному центрированному процессу $X(t)$, реализации которого с вероятностью 1 имеют бесконечную энергию и, следовательно, не имеют преобразования Фурье. Определим усеченный процесс следующим образом:

$$X_T(t) = \begin{cases} X(t) & \text{при} \quad |t| \leqslant \frac{T}{2}; \\ 0 & \text{при} \quad |t| > \frac{T}{2}. \end{cases} \tag{2.37}$$

Усеченный процесс имеет конечную энергию, а следовательно, и спектральную плотность энергии

$$M\left\{\left|\int_{-T/2}^{T/2} X(t)\, e^{-i\,2\pi ft}\, dt\right|^2\right\}. \tag{2.38}$$

Разделив спектральную плотность энергии процесса $X_T(t)$ на его длительность T, получим спектральную плотность его мощности

$$G_T(f) = M\left\{\frac{1}{T}\left|\int_{-T/2}^{T/2} X(t)\, e^{-i\,2\pi ft}\, dt\right|^2\right\}. \tag{2.39}$$

Теперь можно устремить интервал T к бесконечности и определить *спектральную плотность мощности* (СПМ) стационарного процесса:

$$G(f) = \lim_{T\to\infty} M\left\{\frac{1}{T}\left|\int_{-T/2}^{T/2} X(t)\, e^{-i\,2\pi ft}\, dt\right|^2\right\}. \tag{2.40}$$

Легко убедиться, что $G(f)$ — неотрицательная четная функция, ограничивающая (вместе с осью абсцисс) площадь, равную мощности

$$P_X = \lim_{T\to\infty} M\left\{\frac{1}{T}\int_{-T/2}^{T/2} X^2(t)\, dt\right\}$$ центрированного процесса $X(t)$.

Напомним, что для центрированных эргодических процессов мощность совпадает с дисперсией.

Во многих случаях удобнее пользоваться односторонней спектральной плотностью мощности, заданной при $f>0$: $G_0(f) = 2G(f)$. Множитель 2 обеспечивает равенство

$$\int_0^\infty G_0(f)\, df = P_X.$$

Нормированной спектральной плотностью мощности $\gamma(f)$ будем называть отношение $G(f)/P_X$. Отметим, что для любых процессов $X(t)$ с конечной или бесконечной энергией можно определить *нормированный энергетический спектр* (НЭС) формулой

$$\gamma(f) = \lim_{T\to\infty} M\left\{\frac{\left|\int\limits_{-T/2}^{T/2} x(t)\, e^{-i2\pi f}\, dt\right|^2}{\int\limits_{-T/2}^{T/2} X^2(t)\, dt}\right\}. \qquad (2.41)$$

Для процессов из L_2 в (2.41) существуют раздельно пределы числителя (2.36) и знаменателя (2.35), и в этом случае $\gamma(f)$ представляет нормированную спектральную плотность энергии. Для стационарных процессов пределы числителя и знаменателя порознь не существуют, но, как легко убедиться, предел их отношения совпадает с $G(f)/P_X$. Таким образом, нормированный энергетический спектр является универсальной характеристикой сигнала. Он характеризует относительное распределение энергии (или мощности) процесса в частотной области. Ее график всегда ограничивает площадь, равную 1. В инженерной практике чаще всего под спектром понимают именно нормированный энергетический спектр. Для примера на рис. 2.2 показаны односторонние нормированные энергетические спектры $\gamma_0(f) = 2\gamma(f)$ (при $f > 0$) для некоторых ансамблей сигналов.

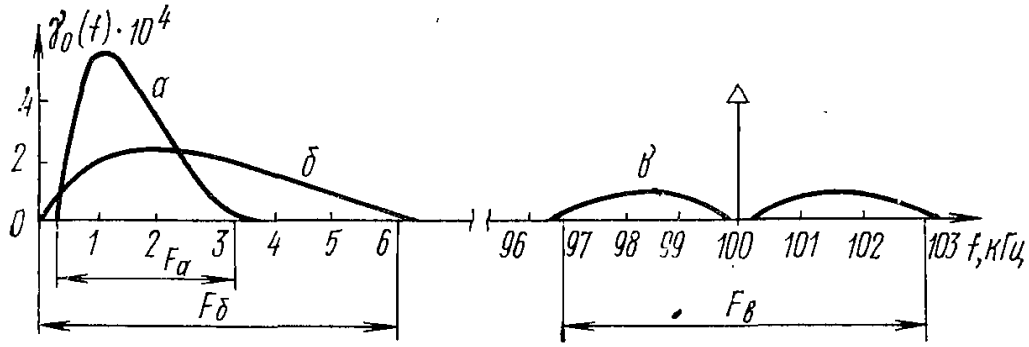


Рис. 2.2. Односторонние нормированные энергетические спектры некоторых сигналов:
а — телефонного; *б* — первичного радиовещательного; *в* — радиотелефонного при амплитудной модуляции (на несущей частоте 100 кГц сосредоточено 70% мощности)

Часто применяют математические модели сигналов, в которых спектр отличен от нуля только на некоторой полосе частот $f_1 \leqslant f \leqslant f_2$, т. е. процессы с *финитным спектром*. Разность $F = f_2 - f_1$ является *шириной спектра*. В реальных условиях жесткого ограничения спектра не бывает и под шириной спектра F понимают ширину минимальной полосы частот, на которой сосредоточена подавляющая часть (обычно 95%) мощности (или энергии) сигнала, как показано на рис. 2.2.

Спектральную плотность мощности стационарного случайного процесса можно определить по его корреляционной функции на основании следующей *теоремы Хинчина—Винера*:

Спектральная плотность мощности $G(f)$ центрированного стационарного случайного процесса является преобразованием Фурье от корреляционной функции $B_X(\tau)$:

$$G_X(f) = \int\limits_{-\infty}^{\infty} B_X(\tau)\, e^{-i2\pi f\tau}\, d\tau. \qquad (2.42)$$

Доказательство этой теоремы можно найти в [12]. Заметим, что преобразование Фурье от $B(\tau)$ существует, строго говоря, если эта корреляционная функция абсолютно интегрируема и это ее свойство используется при доказательстве теоремы. Однако теорему Хинчина — Винера можно успешно применять и в тех случаях, когда $B(\tau)$ не удовлетворяет свойству абсолютной интегрируемости, если допустить, что $B(\tau)$ и $G(f)$ выражаются через обобщенные функции. Так, например, ФК процесса (2.6), если Ф имеет равномерное распределение на интервале $(0; 2\pi)$, равна $B(\tau)=0{,}5A^2\cos\omega\tau$ и не является абсолютно интегрируемой. Однако ее преобразование Фурье $G(f)=0{,}25A^2[\delta(2\pi f-\omega)+\delta(2\pi f+\omega)]$ имеет физический смысл. Это линейчатый спектр, сосредоточенный на частотах $f=\pm\omega/2\pi$.

Отметим несколько следствий теоремы Хинчина—Винера.

Следствие 1. Из теории преобразования Фурье следует справедливость обратного преобразования

$$B_X(\tau)=\int_{-\infty}^{\infty}G(f)\,e^{i2\pi f\tau}\,df. \tag{2.43}$$

Следствие 2. Значение $G(f)$ на нулевой частоте равно интегралу корреляционной функции. Для доказательства достаточно положить в (2.42) $f=0$.

Следствие 3. Так как $G(f)$ — неотрицательная функция, то из (2.42) следует, что корреляционными функциями случайных процессов могут быть только такие, которые имеют положительное преобразование Фурье. В частности, например, функция

$$f(\tau)=\begin{cases}1 & \text{при } |\tau|<\tau_1;\\ 0 & \text{при } |\tau|>\tau_1,\end{cases}$$

изображаемая на графике прямоугольником, не может быть корреляционной. Функции, имеющие положительное преобразование Фурье, называются *положительно определенными.*

Формулы (2.42) и (2.43) можно записать более компактно, если учесть, что $B_X(\tau)$, так же как и $G(f)$, являются четными функциями:

$$G(f)=2\int_0^{\infty}B_X(\tau)\cos 2\pi f\tau\,d\tau; \tag{2.42a}$$

$$B_X(\tau)=2\int_0^{\infty}G(f)\cos 2\pi f\tau\,df=\int_0^{\infty}G_0(f)\cos 2\pi f\tau\,df. \tag{2.43a}$$

Совершенно аналогично можно получить связь между нормированной функцией корреляции $R_X(\tau)$ и нормированным энергетическим спектром $\gamma(f)$:

$$\left.\begin{aligned}\gamma(f)&=2\int_0^{\infty}R_X(\tau)\cos 2\pi f\tau\,d\tau;\\ R_X(\tau)&=2\int_0^{\infty}\gamma(f)\cos 2\pi f\tau\,df\end{aligned}\right\} \tag{2.44}$$

или для одностороннего спектра

$$\left.\begin{aligned}\gamma_0(f) &= 4\int_0^\infty R_X(\tau)\cos 2\pi f\tau\, d\tau;\\ R_X(\tau) &= \int_0^\infty \gamma_0(f)\cos 2\pi f\tau\, df.\end{aligned}\right\} \qquad (2.44а)$$

Рассмотрим несколько полезных примеров.

1. Найдем ФК и СПМ «случайного синхронного телеграфного сигнала». Под этим понимается центрированный случайный процесс, принимающий с равной вероятностью значения $+1$ и -1, причем смена значения может происходить только в моменты времени, разделенные промежутком T_0 (тактовым интервалом). Значения на разных тактовых интервалах независимы. Пример реализации такого процесса приведен на рис. 2.3. Заметим, что границы тактовых интервалов у разных реализаций не совпадают, так что любой момент времени на интервале $0-T$ может с равной вероятностью оказаться моментом начала такта.

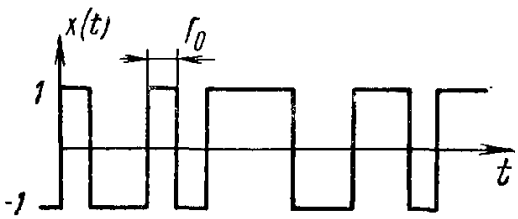


Рис. 2.3. Реализация синхронного телеграфного сигнала

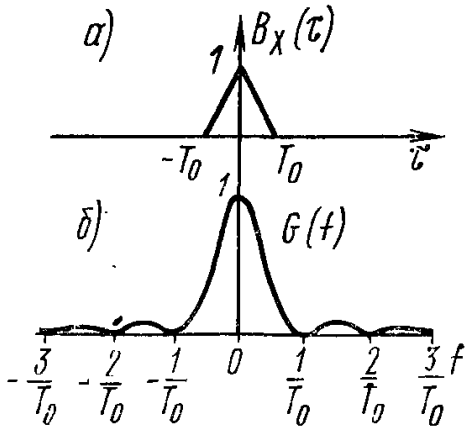


Рис. 2.4. Функция корреляции *(а)* и спектральная плотность мощности *(б)* синхронного телеграфного сигнала

Для определения ФК рассмотрим два сечения в моменты t_1 и t_2, обозначим t_2-t_1 через τ и найдем математическое ожидание произведения $X(t_1)X(t_1+\tau)$. Если $\tau>T_0$, то эти сечения принадлежат разным тактовым интервалам и наше произведение может с равной вероятностью принимать значения $+1$ и -1, так что его математическое ожидание равно нулю. Если же $\tau<T_0$, то возможны два случая: случай A, когда они принадлежат одному интервалу и, следовательно, $X(t_1)\,X(t_1+\tau)=1$, и случай B, когда они принадлежат разным интервалам и $X(t_1)X(t_1+\tau)$ может с равной вероятностью равняться $+1$ и -1. Поэтому при $\tau<T_0$ математическое ожидание $X(t_1)X(t_1+\tau)$ равно вероятности $P(A)$ того, что оба сечения оказались в одном интервале. Легко сообразить, что случай A имеет место, если первое из двух сечений отстоит от начала тактового интервала не более чем на $T_0-|\tau|$, а вероятность этого равна $1-|\tau|/T_0$. Итак,

$$B_X(\tau)=\begin{cases}1-|\tau|/T_0 & |\tau|\leqslant T_0;\\ 0 & |\tau|>T_0.\end{cases}$$

Так как B_X зависит только от разности $t_2-t_1=\tau$, а $\overline{X(t)}=0$, то процесс стационарный. График функции корреляции представлен на рис. 2.4*а*.

Спектральную плотность мощности синхронного телеграфного сигнала вычислим по формуле (2.42а).

$$G(f) = 2\int_0^{\infty} B_X(\tau)\cos 2\pi f\tau\, d\tau = 2\int_0^{T_0}\left(1-\frac{\tau}{T_0}\right)\cos 2\pi f\tau\, d\tau =$$

$$= 2\int_0^{T_0}\cos 2\pi f\tau\, d\tau - \frac{2}{T_0}\int_0^{T_0}\tau\cos 2\pi f\tau\, d\tau = 2\int_0^{T_0}\cos 2\pi f\tau\, d\tau -$$

$$- \frac{2}{T_0}\frac{\tau\sin 2\pi f\tau}{2\pi f}\Bigg|_0^{T_0} + \frac{2}{T_0}\int_0^{T_0}\frac{\sin 2\pi f\tau}{2\pi f}\,d\tau = \frac{2\sin 2\pi f T_0}{2\pi f} - \frac{2\sin 2\pi f T_0}{2\pi f} +$$

$$+ \frac{T_0}{2(\pi f T_0)^2}(1-\cos 2\pi f T_0) = T_0\frac{\sin^2(\pi f T_0)}{(\pi f T_0)^2}.$$

Полученная спектральная плотность показана на рис. 2.4*б*. Заметим, что дисперсия рассмотренного сигнала равна 1, поэтому $B_X(\tau)$ совпадает с $R_X(\tau)$, а $G(f)$ — с $\gamma(f)$.

2. Найдем нормированный энергетический спектр марковского нормального процесса, для которого коэффициент корреляции выражается формулой (2.31). Используя (2.44), находим $\gamma(f) = 2\int_0^{\infty} e^{-\alpha\tau}\cos 2\pi f\tau d\tau = 2\alpha/(\alpha^2+(2\pi f)^2)$. На рис. 2.5 изображены графики функций $R(\tau)$ и $\gamma(f)$.

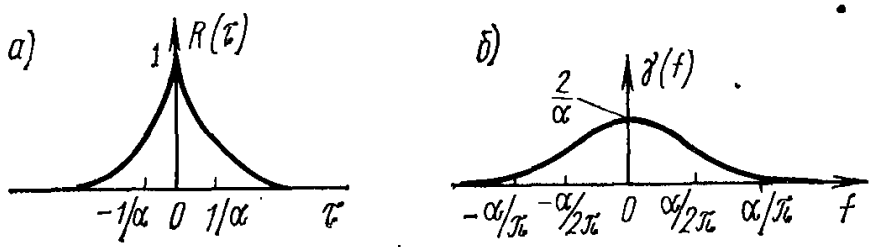


Рис. 2.5. Коэффициент корреляции *(а)* и нормированный энергетический спектр *(б)* марковского гауссовского процесса

3. Пусть $X(t)$ — центрированный стационарный случайный процесс со спектральной плотностью мощности $G_X(f)$. Найдем спектральную плотность мощности $G_Y(f)$ процесса $Y(t) = X(t)\cos(2\pi f_0 t + \Phi)$, где Φ — случайная величина, не зависящая от X и равномерно распределенная на интервале от 0 до 2π. Такая задача часто встречается при изучении модуляции, когда f_0 — несущая частота, а $X(t)$ — модулирующий первичный сигнал.

Легко убедиться, что процесс $Y(t)$ также центрированный. Начнем с отыскания функции корреляции $B_Y(t_1, t_2)$:

$$B_Y(t_1, t_2) = M\{X(t_1)\cos(2\pi f_0 t_1 + \Phi)\, X(t_2)\cos(2\pi f_0 t_2 + \Phi)\} =$$

$$= M\left\{X(t_1)\,X(t_2)\cdot\frac{1}{2}[\cos(2\pi f_0(t_2+t_1)+2\Phi) + \cos(2\pi f_0(t_2-t_1))]\right\}.$$

Так как математическое ожидание произведения независимых величин равно произведению их математических ожиданий, получим

$$B_Y(t_1, t_2) = \frac{1}{2}M\{X(t_1)\,X(t_2)\}\,[M\{\cos[2\pi f_0(t_1+t_2)+2\Phi]\} +$$

$$+ M\{\cos[2\pi f_0(t_2-t_1)]\}].$$

$$\text{Но } M\{\cos[2\pi f_0(t_2+t_1)+\Phi]\}=\frac{1}{2\pi}\int_0^{2\pi}\cos[2\pi f(t_2+t_1)+\Phi]\,d\Phi=0.$$

Окончательно: $B_Y(t_1, t_2)=\frac{1}{2}B_X(t_1, t_2)\cos 2\pi f_0(t_2-t_1)$.

Поскольку $X(t)$ — стационарный процесс и $B_X(t_1, t_2)=B_X(\tau)$, где $\tau=t_2-t_1$, то и $Y(t)$ оказывается стационарным и

$$B_Y(\tau)=\frac{1}{2}B_X(\tau)\cos 2\pi f_0\tau. \tag{2.45}$$

Его спектральная плотность мощности (двусторонняя)

$$G_Y(f)=2\int_0^{\infty}\frac{1}{2}B_X(\tau)\cos 2\pi f_0\tau\cos 2\pi f\tau\,d\tau=$$

$$=\frac{1}{2}\int_0^{\infty}B_X(\tau)[\cos 2\pi(f+f_0)\tau+\cos 2\pi(f-f_0)\tau]\,d\tau=$$

$$=\frac{1}{4}G_X(f+f_0)+\frac{1}{4}G_X(f-f_0). \tag{2.46}$$

Иллюстрацией полученного результата является рис. 2.6. Спектр $G_Y(f)$ состоит из двух боковых полос.

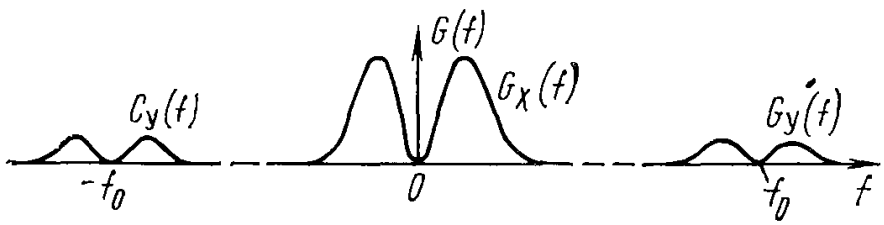


Рис. 2.6. Преобразование спектральной плотности при умножении процесса на гармонический сигнал

Следует отметить, что если бы величина Φ не была случайной или ее распределение вероятностей было несимметричным, то процесс $Y(t)$ не был стационарным.

4. Стационарный процесс $W(t)$ с равномерной спектральной плотностью мощности в некоторой полосе частот называют *квазибелым шумом*, по аналогии с белым светом, т. е. электромагнитными волнами, имеющими равномерный спектр в области видимых частот. Пусть

$$G_W(f)=\begin{cases}N=\text{const} & \text{при } |f|\leqslant F;\\ 0 & \text{при } |f|>F.\end{cases} \tag{2.47}$$

Найдем его корреляционную функцию (рис. 2.7):

$$B_W(\tau)=2\int_0^{F}N\cos 2\pi f\tau\,df=2NF\,\frac{\sin 2\pi F\tau}{2\pi F\tau}. \tag{2.48}$$

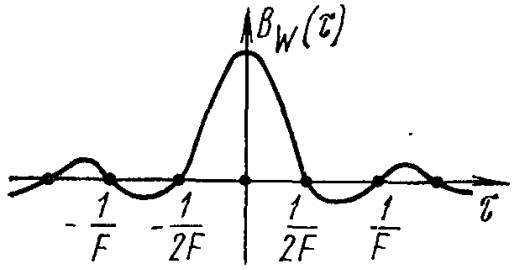


Рис. 2.7. Функция корреляции низкочастотного квазибелого шума

Следует обратить внимание на то, что при значениях τ, кратных $1/(2F)$, $B_W(\tau)=0$. Таким образом, сечения процесса, разделенные интервалом $k/(2F)$ (где k — целое число), не коррелированы между собой.

Если беспредельно увеличивать граничную частоту F, то придем к процессу, у которого любые два несовпадающих сечения не коррелированы. Такой процесс называется *белым шумом*. Его спектральная плотность N постоянна на всех частотах, а ФК равна

$$B_W(\tau)=N\delta(\tau)=\frac{N_0}{2}\delta(\tau), \qquad (2.49)$$

где $N_0=2N$ — односторонняя спектральная плотность белого шума, которую называют также интенсивностью шума.

Белый шум представляет собой не реальный физический процесс, а математическую идеализацию, весьма полезную и широко применяемую. Его дисперсия $D\{W(t)\}=B_W(0)=\infty$. На практике приходится часто встречаться с процессами, имеющими равномерную спектральную плотность в весьма широкой полосе частот, более широкой, чем полосы пропускания цепей, на которые они воздействуют. Типичным примером является тепловой шум (1.9), имеющий спектральную плотность, которая практически равномерна для частот $f\ll(kT/h)$, что при температуре $T=300$ К составляет $6\cdot10^{12}$ Гц. В большинстве случаев замена такого шума идеализированным белым не приводит к погрешности. Белый шум часто используют для определения других случайных процессов. Так, марковский процесс можно задать уравнением (2.16), где $N(t)$ — белый шум.

Если выделить из белого шума полосу частот (f_1, f_2) (например, пропустив его через идеальный полосовой фильтр), то мощность шума на выходе будет, очевидно, равна $N_0(f_2-f_1)$.

В некоторых случаях удобно пользоваться идеализированным нестационарным процессом с ФК $B_W(t,\ t+\tau)=\frac{1}{2}N_0(t)\delta(\tau)$, где $N_0(t)$ — некоторая функция времени. Такой процесс называют нестационарным белым шумом.

5. Найдем функцию корреляции высокочастотного квазибелого шума, у которого

$$G_W(f)=\begin{cases}N=\text{const} & \text{при } f_1<|f|<f_2;\\ 0 & \text{при прочих значениях } f.\end{cases}$$

Для него

$$B_W(\tau)=2\int_{f_1}^{f_2}N\cos 2\pi f\tau\,d\tau=2N\frac{\sin 2\pi f_2\tau-\sin 2\pi f_1\tau}{2\pi\tau}=$$

$$= N_0 \frac{\sin \pi (f_2 - f_1)\tau}{\pi \tau} \cos \pi (f_2 + f_1)\tau = N_0 F \frac{\sin \pi F \tau}{\pi F \tau} \cos 2\pi f_0 \tau, \qquad (2.50)$$

где $N_0 = 2N$; $F = f_2 - f_1$; $f_0 = (f_1 + f_2)/2$ (рис. 2.8).

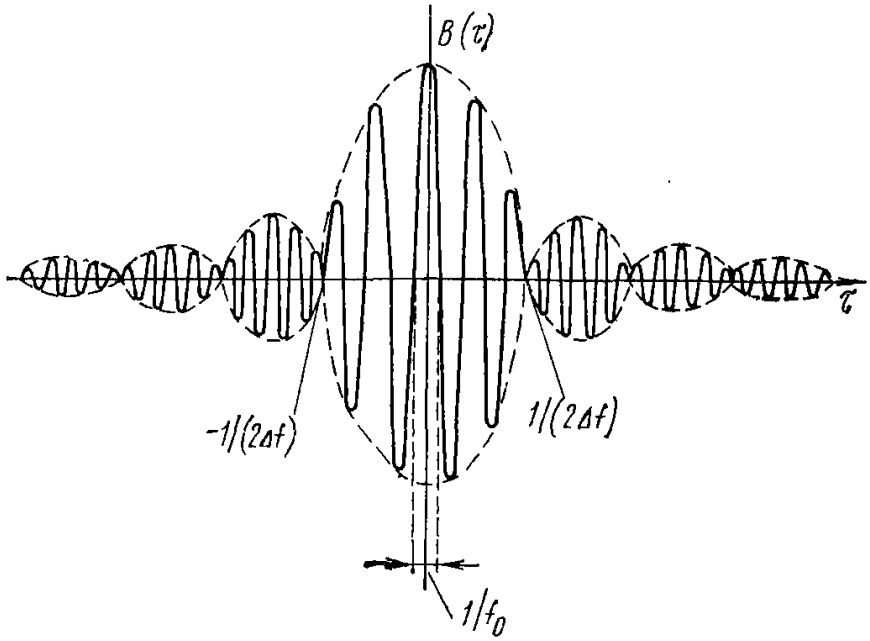


Рис. 2.8. Функция корреляции высокочастотного квазибелого шума ($2\Delta f = F$)

Заметим, что полученная корреляционная функция имеет осциллирующий характер благодаря множителю $\cos 2\pi f_0 \tau$. Это свойственно всем процессам с относительно узкополосным спектром, ширина которого F значительно меньше средней частоты $f_{ср}$. Аналогичный характер функции корреляций был получен в примере 3 [см. (2.45)]. Из (2.50) видно, что сечения высокочастотного квазибелого шума не коррелированы между собой, если интервал между ними τ кратен $1/F$, а также если $\tau = (2n+1)/4f_0$, где n — целое число.

Рассматривая приведенные выше примеры, нетрудно усмотреть, что увеличение интервала корреляции приводит к сокращению ширины спектра и наоборот. Это является проявлением общего свойства любой пары функций, связанных между собой преобразованием Фурье.

2.3. КОМПЛЕКСНОЕ И КВАЗИГАРМОНИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ СИГНАЛОВ

При изучении гармонических сигналов широко пользуются символическим методом, заменяя действительный сигнал $a \cos(\omega_0 t + \varphi) = \mathrm{Re}\{a\, e^{i(\omega_0 t + \varphi)}\}$ комплексной функцией $a\, e^{i(\omega_0 t + \varphi)}$. Это облегчает решение многих задач и приводит к правильным результатам, если полученное комплексное решение заменить его действительной частью.

Обобщением символического метода является представление любого сложного сигнала $x(t)$ (в частности, реализации случайного процесса) комплексной функцией

$$\dot{x}(t) = a(t)\, e^{i\theta(t)} = x(t) + i\tilde{x}(t), \tag{2.51}$$

где

$$x(t) = \operatorname{Re}\dot{x}(t) = a(t)\cos\theta(t), \quad \tilde{x}(t) = \operatorname{Im}\dot{x}(t) = a(t)\sin\theta(t). \tag{2.52}$$

В частном случае гармонического сигнала $a(t) = a = \text{const}$ и $\theta(t) = \omega_0 t + \varphi$. При этом *сопряженный сигнал* $\tilde{x}(t) = a\sin(\omega_0 t + \varphi)$ отличается от действительного сигнала только сдвигом фазы на $-\pi/2$. Естественно и в общем случае негармонического сигнала $x(t)$ определить сопряженный сигнал как результат поворота фаз всех гармонических составляющих на $-\pi/2$. Другими словами, если $x(t)$ можно представить интегралом Фурье

$$x(t) = \frac{1}{2\pi}\int_0^\infty [\alpha(\omega)\cos\omega t + \beta(\omega)\sin\omega t]\, d\omega, \tag{2.53}$$

то

$$\tilde{x}(t) = \frac{1}{2\pi}\int_0^\infty [\alpha(\omega)\sin\omega t - \beta(\omega)\cos\omega t]\, d\omega. \tag{2.54}$$

Можно показать, что два сигнала $x(t)$ и $\tilde{x}(t)$, удовлетворяющие этим условиям, связаны между собой следующими линейными интегральными преобразованиями, называемыми *преобразованиями Гильберта*:

$$\tilde{x}(t) = \frac{1}{\pi}\int_0^\infty \frac{x(t-\tau) - x(t+\tau)}{\tau}\, d\tau,$$

$$x(t) = -\frac{1}{\pi}\int_0^\infty \frac{\tilde{x}(t-\tau) - \tilde{x}(t+\tau)}{\tau}\, d\tau. \tag{2.55}$$

Комплексный сигнал $x(t) + i\tilde{x}(t)$, если $\tilde{x}(t)$ определено по (2.55), называют *аналитическим* (или гильбертовским) *сигналом*.

Поскольку фазы всех составляющих спектральной плотности $S_x(i\omega)$ сопряженного сигнала $\tilde{x}(t)$ в области положительных частот сдвинуты на $-\pi/2$ относительно $S_x(i\omega)$, а умножение на i означает сдвиг фаз на $+\pi/2$, $i\tilde{x}(t)$ имеет в области положительных частот такой же спектр, как и $x(t)$. В области же отрицательных частот преобразование Гильберта соответствует повороту фаз на $+\pi/2$, и поэтому в спектре $i\tilde{x}(t)$ фазы отрицательных частот сдвинуты на π относительно спектра $x(t)$. В результате спектр аналитического сигнала в области отрицательных частот тождественно равен нулю, тогда как в области положительных частот спектры $x(t)$ и $i\tilde{x}(t)$ складываются в одной фазе:

$$S_x(i\omega) = \begin{cases} 2S_x(i\omega) & \omega > 0 \\ 0 & \omega < 0 \end{cases}. \tag{2.56}$$

Все сказанное можно распространить и на случайные процессы. Ансамбль комплексных функций действительной переменной t, $\dot{X}(t)=X(t)+\mathrm{i}\,Y(t)$ с заданными n-мерными распределениями вероятностей представляет собой комплексный случайный процесс, который можно рассматривать как совокупность двух действительных процессов, $X(t)$ и $Y(t)$.

Математическое ожидание комплексного случайного процесса

$$M\{\dot{X}(t)\}=M\{X(t)\}+\mathrm{i}\,M\{Y(t)\} \tag{2.57}$$

представляет собой в общем случае комплексную функцию времени. Его дисперсия $D\{\dot{X}(t)\}=M\{|\mathring{\dot{X}}(t)|^2\}$ также в общем случае является функцией времени, но не комплексной, а действительной и к тому же неотрицательной. Здесь $\mathring{\dot{X}}(t)=\dot{X}(t)-M\{\dot{X}(t)\}$ — центрированный процесс.

Функция корреляции комплексного процесса определяется следующим образом:

$$B_{\dot{X}}(t_1,\ t_2)=M\{\mathring{\dot{X}}^*(t_1)\,\mathring{\dot{X}}(t_2)\}, \tag{2.58}$$

где * означает комплексно-сопряженную величину. Она является в общем случае комплексной функцией двух действительных аргументов. При этом $B_{\dot{X}}(t_2,\ t_1)=B^*_{\dot{X}}(t_1,\ t_2)$. Легко убедиться, что $B_{\dot{X}}(t_1,\ t_1)=D\{\dot{X}(t_1)\}$.

Комплексный случайный процесс стационарен (в широком смысле), если его математическое ожидание не зависит от времени, а функция корреляции зависит только от $\tau=t_2-t_1$. Из этого следует, что и дисперсия стационарного комплексного процесса не зависит от времени.

В дальнейшем будем говорить только об *аналитических* комплексных случайных процессах, в которых $Y(t)=\tilde{X}(t)$:

$$\tilde{X}(t)=\frac{1}{\pi}\int_0^\infty \frac{X(t-\tau)-X(t+\tau)}{\tau}\,d\tau. \tag{2.59}$$

Аналитические случайные процессы обладают свойствами, аналогичными детерминированным аналитическим сигналам. В частности, если действительный случайный сигнал $X(t)$ имеет нормированный (двусторонний) энергетический спектр $\gamma_X(f)$, то у аналитического сигнала НЭС равен

$$\gamma_{\dot{X}}(f)=\begin{cases} 2\gamma_X(f) & f>0 \\ 0 & f<0 \end{cases}. \tag{2.60}$$

Таким образом, НЭС аналитического сигнала совпадает с односторонним НЭС его действительной части.

Важным параметром аналитического сигнала является функция взаимной корреляции между $X(t)$ и $\tilde{X}(t)$:

$$B_{X\tilde{X}}(t_1,\ t_2)=M\{\mathring{X}(t_1)\,\mathring{\tilde{X}}(t_2)\}. \tag{2.61}$$

Для стационарного аналитического сигнала она зависит только от разности $t_2-t_1=\tau$, причем при $\tau=0$ она равна нулю, в чем легко убедиться из (2.59):

$$B_{X\hat{X}}(0)=M\{\overset{\circ}{X}(t)\,\widetilde{\overset{\circ}{X}}(t)\}=M\left\{\overset{\circ}{X}(t)\frac{1}{\pi}\int_0^\infty\frac{\overset{\circ}{X}(t-\tau)-\overset{\circ}{X}(t+\tau)}{\tau}\,d\tau\right\}=$$

$$=\frac{1}{\pi}M\left\{\int_0^\infty\frac{\overset{\circ}{X}(t)\overset{\circ}{X}(t-\tau)-\overset{\circ}{X}(t)\overset{\circ}{X}(t+\tau)}{\tau}\,d\tau\right\}=$$

$$=\frac{1}{\pi}\int_0^\infty\frac{B_X(-\tau)-B_X(\tau)}{\tau}\,d\tau=0, \qquad (2.62)$$

так как для стационарного процесса $B_X(-\tau)=B_X(\tau)$. Таким образом, функции $X(t)$ и $\widetilde{X}(t)$ не коррелированы в совпадающие моменты времени.

Представим аналитический случайный сигнал в экспоненциальной форме:

$$\dot{X}(t)=A(t)\,e^{i\Psi(t)}, \qquad (2.63)$$

где

$$A(t)=|\dot{X}(t)|=\sqrt{X^2(t)+\widetilde{X}^2(t)} \qquad (2.64)$$

— действительный неотрицательный случайный процесс, называемый огибающей сигнала $X(t)$; $\Psi(t)$ — действительный случайный процесс такой, что

$$\sin\Psi(t)=\widetilde{X}(t)/A(t),\quad \cos\Psi(t)=X(t)/A(t) \qquad (2.65)$$

(фаза сигнала).

Очевидно, что огибающая и фаза полностью определяются (в вероятностном смысле) действительным сигналом $X(t)$. Из определения огибающей видно, что $A(t)\geqslant|X(t)|$, а в тех точках, где имеет место равенство $A(t)=X(t)$, сигнал и его огибающая имеют одинаковые производные[1] $dX/dt=dA/dt$.

Аналитический случайный сигнал можно представить случайным векторным процессом на комплексной плоскости. На рис. 2.9 показана одна из реализаций вектора $\dot{X}(t)$ в момент $t=0$. Длина этого вектора равна $A(0)$, а угол между ним и действительной осью составляет $\Psi(0)$. С течением времени конец век-

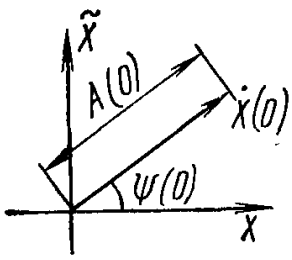


Рис. 2.9. Представление аналитического сигнала вращающимся вектором

[1] Аналогично в тех точках, где $A(t)=-X(t)$, имеет место равенство $\frac{dX}{dt}=\frac{d}{dt}[-A(t)]$.

тора перемещается по некоторой траектории, так что в любой момент его длина равна $A(t)$ и он образует угол $\Psi(t)$ с действительной осью.

Угловая скорость вектора $\dot{X}(t)$, $\Omega(t)=d\Psi/dt$ называется *мгновенной частотой сигнала*. Ее можно выразить через действительный сигнал $X(t)$ и сопряженный с ним сигнал $\widetilde{X}(t)$:

$$\Omega(t)=\frac{\widetilde{X}'(t)\,X(t)-X'(t)\,\widetilde{X}(t)}{X^2(t)+\widetilde{X}^2(t)}, \tag{2.66}$$

где штрихи обозначают производные по времени. В частности, как легко убедиться, для гармонического сигнала $x(t)=a\cos(\omega_1 t+\varphi)=\mathrm{Re}[a\,e^{j(\omega_1 t+\varphi)}]$ мгновенная частота совпадает с ω_1.

Выберем некоторую произвольную частоту ω_0 и обозначим $\Phi(t)=\Psi(t)-\omega_0 t$. Тогда

$$\dot{X}(t)=A(t)\,e^{j[\omega_0 t+\Phi(t)]}. \tag{2.67}$$

Случайная функция $\Phi(t)$ называется *мгновенной начальной фазой* относительно частоты ω_0. Очевидно, что она зависит от выбора ω_0.

Действительный сигнал и сопряженный с ним можно представить в *квазигармонической* форме, непосредственно вытекающей из (2.67):

$$\left.\begin{aligned}X(t)&=A(t)\cos[\omega_0 t+\Phi(t)],\\ \widetilde{X}(t)&=A(t)\sin[\omega_0 t+\Phi(t)].\end{aligned}\right\} \tag{2.68}$$

Приведем без доказательства одно свойство аналитического сигнала. Если все гармонические составляющие всех реализаций процесса $X(t)$, выраженного формулой (2.68), сдвинуть по фазе на одинаковый угол θ, то полученный при этом процесс, который обозначим $X_\theta(t)$, можно представить в квазигармонической форме $X_\theta(t)=A(t)\cos[\omega_0 t+\Phi(t)+\theta]$. Отсюда с помощью простых тригонометрических преобразований с учетом (2.68) получим следующее представление $X_\theta(t)$:

$$X_\theta(t)=X(t)\cos\theta-\widetilde{X}(t)\sin\theta. \tag{2.69}$$

2.4. ОГИБАЮЩАЯ И ФАЗА СЛУЧАЙНОГО ПРОЦЕССА

Квазигармоническое представление возможно для любого сигнала, однако наиболее полезно оно при изучении относительно *узкополосных* сигналов, у которых ширина спектра F значительно меньше средней частоты $f_{ср}$. Для таких сигналов огибающая $A(t)$ является низкочастотным процессом, т. е. изменяется медленно по сравнению с изменениями $X(t)$. Это же верно и для мгновенной начальной фазы $\Phi(t)$ и для ее производной $d\Phi/dt=\Omega(t)-\omega_0$ при условии, что частота ω_0 выбрана вблизи средней круговой частоты спектра $2\pi f_{ср}$. При наблюдении на осциллографе такой процесс имеет вид колебания, промодулированного по амплитуде и фазе.

Выражение (2.67) можно переписать также следующим образом:

$$\dot{X}(t) = \dot{A}(t)\, e^{j\omega_0 t}, \tag{2.70}$$

где $\dot{A}(t) = A(t) e^{j\Phi(t)}$ — *комплексная огибающая* сигнала. В отличие от действительной огибающей $A(t)$, комплексная огибающая не определяется однозначно сигналом $X(t)$, а зависит также от выбора произвольной частоты ω_0. Из очевидного равенства

$$\dot{A}(t) = \dot{X}(t)\, e^{-j\omega_0 t} \tag{2.70а}$$

следует, что комплексную огибающую можно рассматривать как результат сдвига спектра аналитического сигнала вниз на ω_0[1].

Введем обозначения для действительной и мнимой частей комплексной огибающей:

$$A_c(t) = \operatorname{Re}\dot{A}(t) = A(t)\cos\Phi(t), \quad A_s(t) = -\operatorname{Im}\dot{A}(t) = -A(t)\sin\Phi(t). \tag{2.71}$$

С помощью этих функций можно получить еще одно полезное представление сигнала $X(t)$:

$$X(t) = A(t)\cos[\omega_0 t + \Phi(t)] = A(t)\cos\Phi(t)\cos\omega_0 t - $$
$$- A(t)\sin\Phi(t)\sin\omega_0 t = A_c(t)\cos\omega_0 t + A_s(t)\sin\omega_0 t. \tag{2.72}$$

Таким образом, сигнал $X(t)$ разложен на две *квадратурные составляющие*: $A_c(t)\cos\omega_0 t$ и $A_s(t)\sin\omega_0 t$.

Функции $A_c(t)$ и $A_s(t)$ зависят от выбора частоты ω_0. Для узкополосного сигнала эти функции, при надлежащем выборе ω_0, являются, как и $A(t)$, медленно меняющимися (низкочастотными). Их, однако, не следует называть огибающими квадратурных составляющих, поскольку, в отличие от $A(t)$, они принимают не только положительные, но и отрицательные значения.

Зная функции распределения исходного процесса $X(t)$, можно обычными методами теории вероятностей найти функции распределения для всех вновь введенных процессов: $\tilde{X}(t)$, $\dot{X}(t)$, $A(t)$, $\Psi(t)$, $\Phi(t)$, $\omega(t)$ и т. д. В общем случае это не простая задача. Поэтому остановимся подробнее только на наиболее важном случае для практики, когда $X(t)$ — гауссовский процесс.

Рассмотрим вначале центрированный стационарный гауссовский процесс $X(t)$. Поскольку преобразование Гильберта линейно, то, учитывая инвариантность нормального распределения при любых линейных операциях, можно утверждать, что $\tilde{X}(t)$ — также гауссовский процесс.

На основании того, что $\tilde{X}(t)$ отличается от $X(t)$ только сдвигом фаз всех составляющих, и вспоминая, что спектральная плотность мощности $G(f)$ не зависит от фазовых соотношений, прихо-

[1] Отсюда, между прочим, видно, что спектр $\dot{A}(t)$ не лежит целиком в области $f > 0$, и, следовательно, комплексный процесс $\dot{A}(t)$ не является аналитическим.

дим к выводу, что $G_{\tilde{X}}(f)=G_X(f)$. Из этого на основании теоремы Хинчина—Винера следует равенство корреляционных функций $B_{\tilde{X}}(\tau)=B_X(\tau)$, а отсюда и равенство дисперсий $D\{\tilde{X}(t)\}=$ $=B_{\tilde{X}}(0)=B_X(0)=D\{X(t)\}$. В совпадающие моменты времени $X(t)$ и $\tilde{X}(t)$ между собой не коррелированы [см. ф-лу (2.62)], а поскольку они гауссовские, то и независимы. Комплексная огибающая $\dot{A}(t)$, полученная из $\dot{X}(t)$ посредством линейной операций (2.70а), обладает теми же свойствами: $A_c(t)$ и $A_s(t)$ также являются центрированными гауссовскими процессами, независимыми в совпадающие моменты времени. Дисперсия аналитического сигнала

$$D\{\dot{X}(t)\}=M\{|\dot{X}(t)|^2\}=M\{X^2(t)+\tilde{X}^2(t)\}=D\{X(t)\}+$$

$$+D\{\tilde{X}(t)\}=2D\{X(t)\}.$$

Такова же, как легко проверить, и дисперсия комплексной огибающей $\dot{A}(t)$. Нетрудно также убедиться в том, что $D\{A_c(t)\}=$ $=D\{A_s(t)\}=D\{X(t)\}$ *.

Совместную плотность вероятности $X(t)$ и $\tilde{X}(t)$ в одном сечении t можно записать на основании (2.12) так:

$$w(x;\ \tilde{x};\ t)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{\tilde{x}^2}{2\sigma^2}\right)=$$

$$=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+\tilde{x}^2}{2\sigma^2}\right). \qquad (2.73)$$

Здесь σ^2 — дисперсия, одинаковая для X и для $\tilde{X}$.

Процессы $A(t)$ и $\Psi(t)$, полученные из $X(t)$ и $\tilde{X}(t)$ с помощью нелинейных операций, стационарны, но конечно не являются гауссовскими. Найдем совместную функцию распределения A и Ψ в некотором сечении t, т. е. функцию

$$F(a,\ \psi)=P\{A(t)\leqslant a;\ \Psi(t)\leqslant\psi\}.$$

Геометрическая интерпретация этой задачи представлена на рис. 2.10, где по оси ординат отложены значения $\tilde{x}(t)$, а по оси абсцисс — значения $x(t)$. Искомая функция распределения является вероятностью того, что конец вектора, представляющего аналитический сигнал $\dot{X}(t)$, находится внутри заштрихованной области V.

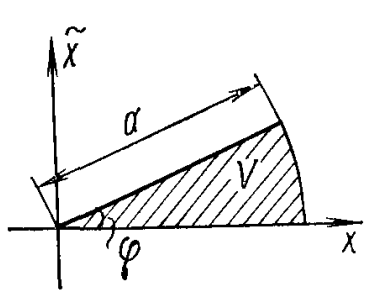


Рис. 2.10. К вычислению распределения вероятностей огибающей и фазы центрированного нормального процесса

Для нахождения функции распределения $F(a, \Psi)$ можно проинтегрировать

* Если спектры $A_c(t)$ и $A_s(t)$ лежат в полосе частот от 0 до F, то в соответствии с (2.46) ширина спектра $X(t)$ равна $2F$. Поэтому средняя спектральная плотность $A_c(t)$ (или $A_s(t)$) в указанной полосе вдвое больше средней спектральной плотности $X(t)$.

(2.73) по области V. Переходя при этом к полярной системе координат ρ, φ по известным формулам [1]

$$x=\rho\cos\varphi;\ \tilde{x}=\rho\sin\varphi;\ dxd\tilde{x}=\rho\,d\rho\,d\varphi$$

получим

$$F(a,\ \psi)=\iint\limits_V \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}[x^2+\tilde{x}^2]\right\}dxd\tilde{x}=$$

$$=\int\limits_0^a \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}\rho^2\right\}\rho\,d\rho\int\limits_0^{\psi} d\varphi=\left[1-\exp\left(-\frac{a^2}{2\sigma^2}\right)\right]\frac{\psi}{2\pi}\ . \qquad (2.74)$$

Таким образом, двумерная функция распределения оказалась произведением двух одномерных функций:

$$F(a)=1-\exp\left(-\frac{a^2}{2\sigma^2}\right)(a\geqslant 0),\ F(\psi)=\frac{\psi}{2\pi}\ (0\leqslant\psi\leqslant 2\pi). \qquad (2.75)$$

Это значит, что огибающая $A(t)$ и фаза $\Psi(t)$ в одном сечении независимы [2].

Плотность распределения огибающей найдем как производную функции распределения:

$$W(a)=\frac{dF(a)}{da}=\frac{a}{\sigma^2}\,e^{-\frac{a^2}{2\sigma^2}},\ (a\geqslant 0). \qquad (2.76)$$

Это известное *распределение Рэлея*, график которого показан на рис. 2.11 вместе с графиком гауссовского распределения мгновенных значений $w(x)$. Наиболее вероятное значение $A(t)$ равно σ. Математическое ожидание $\overline{A(t)}=\sigma\sqrt{\pi/2}$. Математическое ожидание квадрата огибающей $\overline{A^2(t)}=2\sigma^2$, откуда дисперсия $D\{A(t)\}=\overline{A^2(t)}-[\overline{A(t)}]^2=\sigma^2(2-\pi/2)$.

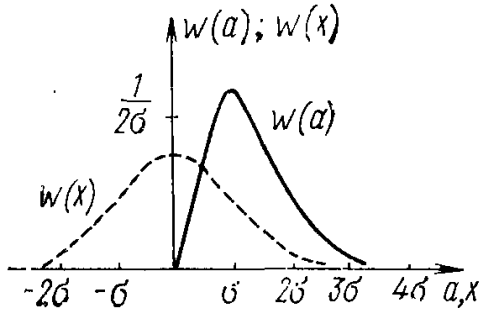


Рис. 2.11. Плотность вероятности мгновенного значения и огибающей нормального центрированного процесса

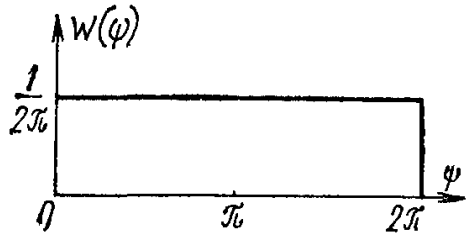


Рис. 2.12. Плотность вероятности фазы нормального центрированного процесса

[1] Заметим, что при этом фаза $\Psi(t)$ определяется с точностью до величины, кратной 2π, иначе говоря, «по модулю 2π». Это не влияет на изучение процесса $X(t)$, поскольку $\Psi(t)$ фигурирует как аргумент косинуса или синуса.

[2] Можно показать, что в центрированном стационарном гауссовском процессе $A(t)$ и $\Psi(t)$ независимы и в разных сечениях.

Для плотности вероятности фазы имеем (рис. 2.12)

$$w(\psi)=\frac{dF(\psi)}{d\psi}=\frac{1}{2\pi}\ (0\leqslant\psi<2\pi). \tag{2.77}$$

Следовательно, фаза стационарного гауссовского центрированного процесса распределена равномерно на интервале протяженностью 2π.

Мгновенная начальная фаза $\Phi(t)$ рассматриваемого процесса также равномерно распределена на интервале протяженностью 2π (если ее определять по модулю 2π). Действительно, можно отыскать совместную функцию распределений $A(t)$ и $\Phi(t)$, исходя из плотности процессов $A_c(t)$ и $A_s(t)$ (2.71). Повторяя буквально те же рассуждения и выкладки, только с заменой X, $\dot{X}$ и ψ на A_c, A_s и Φ, найдем, как и следовало ожидать, плотность (2.76) для $A(t)$ и равномерную плотность вероятностей для $\Phi(t)$.

Воспользуемся этим методом для рассмотрения несколько более сложного случая суммы стационарного центрированного гауссовского процесса и гармонического сигнала:

$$X(t)=\overset{\circ}{X}(t)+u\cos\omega_c t, \tag{2.78}$$

где $\overset{\circ}{X}(t)$ — центрированный стационарный гауссовский процесс; u и ω_c — некоторые постоянные. Этот случай нередко встречается на практике, когда $u\cos\omega_c t$ представляет регулярный сигнал, а $\overset{\circ}{X}(t)$ — флуктуационный шум. Заметим, что $X(t)$ — нестационарный процесс, так как его математическое ожидание $u\cos\omega_c t$ зависит от времени.

Представив $X(t)$ в форме (2.72) и выбирая $\omega_0=\omega_c$, можем записать:

$$\overset{\circ}{X}(t)=\overset{\circ}{A}_c(t)\cos\omega_c t+\overset{\circ}{A}_s(t)\sin\omega_c t;$$

$$X(t)=\overset{\circ}{A}_c(t)\cos\omega_c t+\overset{\circ}{A}_s(t)\sin\omega_c t+u\cos\omega_c t=$$

$$=A_c(t)\cos\omega_c t+A_s(t)\sin\omega_c t=A(t)\cos[\omega_c t+\Phi(t)], \tag{2.79}$$

где $A_s(t)=-A(t)\sin\Phi(t)=\overset{\circ}{A}_s(t)$; $A_c(t)=A(t)\cos\Phi(t)=\overset{\circ}{A}_c(t)+u$; $A(t)=\sqrt{A^2_c(t)+A^2_s(t)}$.

Отсюда видно, что $A_c(t)$ и $A_s(t)$ являются стационарными гауссовскими процессами с одинаковыми дисперсиями σ^2, взаимно некоррелированными в совпадающие моменты времени. При этом $M\{A_s(t)\}=0$, а $M\{A_c(t)\}=u$.

Для нахождения функции распределения

$$F(a,\ \varphi)=P\{A(t)\leqslant a;\ \ \Phi(t)\leqslant\varphi\} \tag{2.80}$$

рассмотрим совместную плотность распределения вероятностей $A_c(t)$ и $A_s(t)$ в некотором сечении t:

$$w(a_c,\ a_s)=\frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}\left[(a_c-u)^2+a_s^2\right]\right\}. \tag{2.81}$$

Функция распределения (2.80) равна интегралу от (2.81) по площади, определяемой (2.80). Переходя к полярным координатам $a_c=\rho\cos\xi$, $a_s=\rho\sin\xi$, $da_c da_s=\rho d\rho d\xi$, находим

$$F(a,\ \varphi)=\int_0^a\int_0^\varphi\frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}[\rho\cos\xi-u)^2+\rho^2\sin^2\xi]\right\}\rho\,d\rho\,d\xi=$$

$$=\frac{1}{2\pi\sigma^2}\int_0^a\int_0^\varphi\rho\exp\left\{-\frac{1}{2\sigma^2}[\rho^2+u^2-2\rho u\cos\xi]\right\}d\rho\,d\xi=$$

$$= \frac{1}{\sigma^2} \exp\left(-\frac{u^2}{2\sigma^2}\right) \int_0^a \rho \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \left[\frac{1}{2\pi} \int_0^{\varphi} \exp\left(\frac{\rho u \cos \xi}{\sigma^2}\right) d\xi\right] d\rho. \qquad (2.82)$$

Не будем вычислять этот интеграл. Заметим лишь, что $F_2(a, \varphi)$ в этом случае не выражается произведением двух функций распределения $F(a)F(\varphi)$, т. е. $A(t)$ и $\Phi(t)$ не являются независимыми. В дальнейшем нам потребуется только распределение вероятностей $A(t)$. Для нахождения $F(a)$ нужно в (2.82) положить $\varphi = 2\pi$, так как по определению $0 \leqslant \varphi < 2\pi$, и вероятность того, что $\varphi \leqslant 2\pi$, равна 1, поэтому $F(a, \varphi)$ при $\varphi = 2\pi$ совпадает с $F(a)$. Итак,

$$F(a) = \frac{1}{\sigma^2} \exp\left(-\frac{u^2}{2\sigma^2}\right) \int_0^a \rho \exp\left(-\frac{\rho^2}{2\sigma^2}\right) \left[\frac{1}{2\pi} \int_0^{2\pi} \exp\left(\frac{\rho u \cos \xi}{\sigma^2}\right) d\xi\right] d\rho.$$

Из теории бесселевых функций известно, что

$$\frac{1}{2\pi} \int_0^{2\pi} e^{x \cos \xi} d\xi = I_0(x), \qquad (2.83)$$

где $I_0(x)$ — модифицированная функция Бесселя нулевого порядка:

$$I_0(x) = J_0(ix) = \sum_{n=0}^{\infty} \frac{1}{(n!)^2} \left(\frac{x}{2}\right)^{2n}. \qquad (2.84)$$

При $x = 0$ эта функция равна единице и монотонно возрастает с x. Таким образом,

$$F(a) = \frac{1}{\sigma^2} \exp\left(-\frac{u^2}{2\sigma^2}\right) \int_0^a \rho \exp\left(-\frac{\rho^2}{2\sigma^2}\right) I_0\left(\frac{\rho u}{\sigma^2}\right) d\rho \qquad (2.85)$$

и плотность вероятностей $A(t)$ для суммы гауссовского шума и гармонического сигнала

$$w(a) = \frac{dF(a)}{da} = \frac{a}{\sigma^2} \exp\left(-\frac{a^2 + u^2}{2\sigma^2}\right) I_0\left(\frac{au}{\sigma^2}\right). \qquad (2.86)$$

Распределение вероятностей, выраженное формулой (2.86), называется *обобщенным распределением Рэлея*, или *распределением Рэлея—Райса*. Очевидно, обычное распределение Рэлея (2.76) является частным случаем распределения (2.86) при $u = 0$.

Заметим без доказательства, что и в более общем случае суммы гауссовского шума с синусоидальным сигналом

$$X(t) = \overset{\circ}{X}(t) + u \cos(\omega_c t + \eta),$$

где η — любая детерминированная или случайная начальная фаза, огибающая $A(t)$, подчиняется распределению (2.86). В частности, если η равномерно распределена на интервале протяженностью 2π, то начальная фаза $\Phi(t)$ также равномерно распреде-

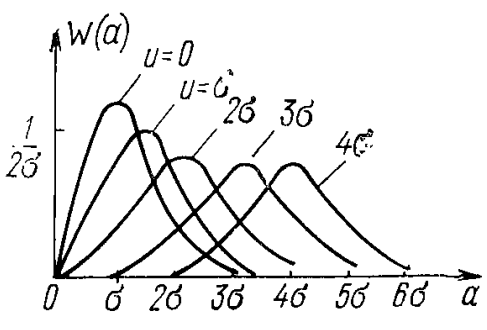


Рис. 2.13. Плотность вероятности огибающей суммы гармонического сигнала и гауссовского шума

лена на том же интервале и не зависит от $A(t)$. Процесс $X(t)$ в этом случае стационарен.

На рис. 2.13 показаны графики плотности (2.86) при разных значениях u. Легко видеть, что чем больше u/σ, т. е. чем больше отношение амплитуды гармонического сигнала к средней квадратичной величине шума, тем больше и наиболее вероятное значение a/σ.

2.5. ПРОСТРАНСТВО СИГНАЛОВ

Подойдем теперь к изучению ансамбля сигналов с другой точки зрения, сделав упор не на вероятностную меру, а на соотношения между отдельными выборочными функциями. Заметим, прежде всего, что обычно два сигнала можно сложить друг с другом и их сумма представляет собой также сигнал (быть может принадлежащий множеству с нулевой вероятностью). Такое сложение происходит, например, когда к полезному сигналу прибавляется аддитивная помеха. Кроме того, сигналы можно умножать на действительные числа, что осуществляется, например, в идеальном усилителе.

В функциональном анализе множество любых элементов x, y, z называется *линейным* (или векторным) *пространством*, если оно удовлетворяет следующим условиям;

1. Для любых двух элементов x и y однозначно определен третий элемент $x+y$, называемый их суммой и также входящий в данное пространство, причем $x+y=y+x$; $x+(y+z)=(x+y)+z$.

2. В линейном пространстве существует нулевой элемент, обозначенный 0, такой, что $x+0=x$ для всех x.

3. Для каждого элемента x линейного пространства существует противоположный ему элемент $(-x)$ такой, что $x+(-x)=0$.

4. Любой элемент пространства можно умножить на любое число из некоторого множества $\{\alpha\}$, называемого множеством скаляров[1], причем αx также принадлежит данному пространству и выполняются следующие соотношения: $1x=x$, $(\alpha+\beta)x=\alpha x+\beta x$, $\alpha(x+y)=\alpha x+\alpha y$, $\alpha(\beta x)=(\alpha\beta)x$.

Примерами линейных пространств могут служить:

множество векторов на плоскости с обычным определением векторного сложения и умножения вектора на число;

множество всех действительных (или всех комплексных) чисел с обычными определениями сложения и умножения;

множество упорядоченных последовательностей из n действительных чисел $\{a_1, a_2, \ldots, a_n\}$, если определить сумму и умножение на скаляр следующим образом:

$$(a_1, a_2, \ldots, a_n)+(b_1, b_2, \ldots, b_n)=(a_1+b_1, a_2+b_2, \ldots, a_n+b_n)$$
$$\alpha(a_1, a_2, \ldots, a_n)=(\alpha a_1, \alpha a_2, \ldots, \alpha a_n) \qquad (2.87)$$

и т. д.

---

[1] Множество $\{\alpha\}$ должно быть полем, т. е. на нем должны быть определены операции сложения и умножения с коммутативными и дистрибутивными свойствами и оно должно содержать нуль и единицу. Примерами полей являются множества всех целых чисел, всех действительных чисел, всех комплексных чисел и др. Существуют и конечные поля (см. ниже).

Во всех приведенных примерах скалярами могут быть любые действительные числа. При другом выборе поля скаляров можно построить, например, такие линейные пространства:

множество всех целых положительных и отрицательных чисел (включая нуль), если поле скаляров содержит тоже только целые числа;

множество, состоящее всего лишь из двух чисел, 0 и 1, если сложение производится по модулю 2, т. е. по правилу

$$1+0=0+1=1, \quad 1+1=0+0=0, \tag{2.88}$$

а скалярами являются также только 0 и 1;

множество упорядоченных последовательностей из n чисел, принимающих значения 0 и 1, при таких же скалярах, причем сложение и умножение производятся поразрядно, т. е. по (2.87), а каждые два числа складываются по модулю 2, т. е. по (2.88). Заметим, что в последних двух пространствах $-x=x$.

Очевидно, что множества, приведенные в последних двух примерах, не представляли бы линейного пространства, если бы сложение определялось не по модулю 2, а обычным образом, так как не выполнялось бы условие 1. Точно так же не являлось бы линейным пространством (при обычном определении сложения) множество всех положительных чисел (не выполняются условия 2 и 3) или множество всех действительных (либо комплексных) чисел, не превосходящих по модулю некоторого числа M (не выполняется условие 1).

Особый интерес для нас представляют *функциональные пространства*, т. е. линейные пространства, элементами которых являются функции. Примером может служить множество всех непрерывных комплекснозначных функций $\overset{\circ}{x}(t)$, заданных на интервале $-T/2 \leqslant t \leqslant T/2$, если скаляры принадлежат полю комплексных чисел. Раздел математики, изучающий функциональные пространства, называется функциональным анализом.

Элементы любого линейного пространства называются *векторами*, их можно рассматривать как обобщение понятия обычных векторов на плоскости или в трехмерном пространстве. Многие свойства обычных векторов переносятся на элементы различных линейных пространств. В частности, ниже будут определены «длина» вектора (которую будем называть *нормой*), угол между двумя векторами, расстояние между векторами и скалярное произведение двух векторов.

Для теории передачи сигналов важным является то обстоятельство, что практически все сигналы (дискретные и непрерывные) и все аддитивные помехи можно рассматривать как векторы в некотором пространстве. Это позволит получить сравнительно просто и наглядно многие важные результаты.

Определим размерность векторного пространства. Особенностью обычных векторов на плоскости (двумерном пространстве) является то, что любой вектор $\mathbf{z}$ можно выразить в виде линейной комбинации любых двух непараллельных между собой векторов $\mathbf{x}$ и $\mathbf{y}$:

$$\mathbf{z}=\alpha_1 \mathbf{x}+\alpha_2 \mathbf{y}. \tag{2.89}$$

Для трехмерного пространства это не имеет места, но зато любой вектор можно выразить линейной комбинацией трех не лежащих в одной плоскости и не параллельных векторов.

Другими словами, в двумерном пространстве три вектора, а в трехмерном пространстве четыре вектора всегда оказываются линейно зависимыми. Распространим это понятие на любое линейное пространство.

Элементы x_1, x_2, ..., x_k линейного пространства называются *линейно зависимыми*, если существуют скаляры α_1, α_2, ..., α_k, не все равные нулю, такие, что

$$\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_k \mathbf{x}_k = \mathbf{0}. \tag{2.90}$$

Очевидно, что в этом случае любой из этих элементов можно выразить линейной комбинацией остальных элементов. Этого нельзя сделать, если (2.90) выполняется только при $\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$. В последнем случае элементы x_1, ..., x_k *линейно независимы.*

Линейное пространство, в котором можно найти n линейно независимых элементов, а любые $n+1$ элементов линейно зависимы, называется *n-мерным.* Если же можно найти произвольное число линейно независимых элементов, то пространство называется *бесконечномерным.*

Базисом в n-мерном пространстве называется любая система из n линейно независимых векторов.

Во многих линейных пространствах можно ввести понятие *скалярного произведения* двух элементов, **x** и **y**, которое обозначается (**x**, **y**). Скалярным произведением называется число (в некоторых пространствах комплексное), удовлетворяющее следующим условиям:

а) $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})^*$ (где * означает комплексную сопряженность);

б) $(\mathbf{x}, \mathbf{x}) \geqslant 0$, т. е. скалярное произведение вектора на самого себя всегда является действительным неотрицательным числом;

в) $(\mathbf{x}, \mathbf{x}) = 0$ только тогда, когда $\mathbf{x} = 0$[1];

г) $(\alpha\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x}, \mathbf{y})$; (2.91)

д) $(\mathbf{x}, \mathbf{y}_1 + \mathbf{y}_2) = (\mathbf{x}, \mathbf{y}_1) + (\mathbf{x}, \mathbf{y}_2)$.

Из условий а) и г) следует, что

$$(\mathbf{x}, \alpha\mathbf{y}) = \alpha^* (\mathbf{x}, \mathbf{y}).$$

Конечномерное пространство со скалярным произведением называется евклидовым.

Нормой вектора **x** называется неотрицательное число, обозначаемое $\|\mathbf{x}\|$ и равное арифметическому значению $\sqrt{(\mathbf{x}, \mathbf{x})}$. Из условия в) следует, что $\|\mathbf{x}\| = 0$ только для нулевого вектора. Из условия г) вытекает важное свойство нормы: $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$. В частности, при $\alpha = -1$ это дает $\|-\mathbf{x}\| = \|\mathbf{x}\|$. Для обычных векторов нормой, как легко убедиться, является длина.

---

[1] Это условие для некоторых представляющих практический интерес пространств не выполняется (см. ниже с. 66). В дальнейшем будем считать, что оно выполнено.

Расстоянием между векторами $\mathbf{x}$ и $\mathbf{y}$, которое обозначается $d(\mathbf{x};\ \mathbf{y})$, называют норму разности этих векторов:

$$d(\mathbf{x};\ \mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| \tag{2.92}$$

Очевидно, что $d(\mathbf{y};\ \mathbf{x}) = \|\mathbf{y}-\mathbf{x}\| = \|-(\mathbf{x}-\mathbf{y})\| = \|\mathbf{x}-\mathbf{y}\| = d(\mathbf{x};\ \mathbf{y})$.

Важную роль в математике играет *неравенство Коши—Буняковского—Шварца*, справедливое для всех линейных пространств:

$$|(\mathbf{x},\ \mathbf{y})| \leqslant \|\mathbf{x}\| \cdot \|\mathbf{y}\|. \tag{2.93}$$

Для его доказательства рассмотрим квадрат нормы элемента $\alpha\mathbf{x}+\mathbf{y}$, который разумеется, неотрицателен:

$$\|\alpha\mathbf{x}+\mathbf{y}\|^2 = (\alpha\mathbf{x}+\mathbf{y},\ \alpha\mathbf{x}+\mathbf{y}) \geqslant 0.$$

Раскрывая скалярное произведение с помощью условий (2.91), найдем

$$(\alpha\mathbf{x}+\mathbf{y},\ \alpha\mathbf{x}+\mathbf{y}) = |\alpha|^2(\mathbf{x},\ \mathbf{x}) + \alpha(\mathbf{x},\ \mathbf{y}) + \alpha^*(\mathbf{y},\ \mathbf{x}) + (\mathbf{y},\ \mathbf{y}) \geqslant 0 \tag{2.94}$$

при всех значениях α. Положим $\alpha = -(\mathbf{x},\ \mathbf{y})^*/(\mathbf{x},\ \mathbf{x})$ и подставим это значение в (2.94):

$$\frac{|(\mathbf{x},\ \mathbf{y})|^2}{\|\mathbf{x}\|^2} - \frac{|(\mathbf{x},\ \mathbf{y})|^2}{\|\mathbf{x}\|^2} - \frac{|(\mathbf{x},\ \mathbf{y})|^2}{\|\mathbf{x}\|^2} + \|\mathbf{y}\|^2 \geqslant 0,$$

откуда непосредственно следует (2.93). Заметим, что в (2.93) имеет место равенство, если $\mathbf{y} = \beta\mathbf{x}$, где β — любой скаляр.

Из неравенства (2.93) вытекают важные свойства нормы и расстояния:

$$\|\mathbf{x}+\mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|,\quad d(\mathbf{x};\ \mathbf{z}) \leqslant d(\mathbf{x},\ \mathbf{y}) + d(\mathbf{y},\ \mathbf{z}). \tag{2.95}$$

Доказательство этих неравенств предоставляем читателю. В применении к обычным векторам на плоскости оба эти неравенства характеризуют известный из элементарной геометрии факт — длина стороны треугольника меньше суммы длин двух других сторон. Поэтому их часто называют «соотношением треугольника».

В пространстве обычных векторов на плоскости скалярное произведение определяется, как известно, следующим образом:

$$(\mathbf{x},\ \mathbf{y}) = \|\mathbf{x}\| \cdot \|\mathbf{y}\| \cdot \cos\varphi,$$

где φ — угол между векторами. Нетрудно убедиться, что при этом выполняются условия (2.91).

По аналогии с этим определим угол φ между элементами $\mathbf{x}$ и $\mathbf{y}$ любого евклидова пространства как

$$\varphi = \arccos\frac{(\mathbf{x},\ \mathbf{y})}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}.$$

Это определение корректно, так как из неравенства (2.93) следует, что $|\cos\varphi| \leqslant 1$.

В частности, если $(\mathbf{x},\ \mathbf{y}) = 0$, то $\varphi = \pm\pi/2$ и элементы пространства $\mathbf{x}$ и $\mathbf{y}$ называются *ортогональными*.

Подмножество векторов $\{\mathbf{x}_k\}$ в евклидовом пространстве называется *ортогональной системой*, если $(\mathbf{x}_k,\ \mathbf{x}_l) = 0$ при $k \neq l$.

Элементы ортогональной системы линейно независимы. Действительно, если в уравнении (2.90) все $\mathbf{x}_i$ взаимно ортогональны, то, умножая обе части уравнения скалярно на $\mathbf{x}_1$, найдем, что $\alpha_1 = 0$. Так как то же самое можно получить для всех α_i, то из этого следует линейная независимость ортогональных векторов. В любом n-мерном пространстве можно построить полный *ортогональный* базис, т. е. систему из n ортогональных векторов.

Ортогональный базис, удовлетворяющий условию

$$(\mathbf{x}_k, \ x_l) = \begin{cases} 0 \text{ при } k \neq l; \\ 1 \text{ при } k = l, \end{cases} \tag{2.96}$$

называется *ортонормированным*. Очевидно, что по любому ортогональному базису $\{\mathbf{x}_k\}$ можно построить ортонормированный, заменив все $\mathbf{x}_k$ на $\mathbf{x}_k / \|\mathbf{x}_k\|$.

2.6. РАЗЛОЖЕНИЕ СИГНАЛОВ В РЯД ПО ОРТОГОНАЛЬНЫМ ФУНКЦИЯМ

Пусть $\{\varphi_k\}$ — полный ортогональный базис, а $\mathbf{x}$ — некоторый вектор в данном пространстве. Определим числа c_k следующим образом:

$$c_k = (\mathbf{x}, \ \varphi_k) \frac{1}{\| \varphi_k \|^2} \tag{2.97}$$

и построим ряд $\sum\limits_k c_k \varphi_k$. В случае конечномерного пространства этот ряд вырождается в конечную сумму. Имеет место равенство

$$\sum_k c_k \varphi_k = \mathbf{x}, \tag{2.98}$$

называемое разложением вектора $\mathbf{x}$ по базису $\{\varphi_k\}$. В случае конечномерного пространства это равенство понимается в обычном смысле, в случае же бесконечномерного пространства — в смысле сходимости по норме, т. е.

$$\lim_{n \to \infty} \left\| \sum_{k=1}^{n} c_k \varphi_k - \mathbf{x} \right\| = 0. \tag{2.99}$$

Ряд (2.98) называют *обобщенным рядом Фурье*, а числа c_k — коэффициентами Фурье по данному базису.

Докажем равенство (2.98) для конечномерного пространства. Предположим, что оно неверно. Тогда вектор $\mathbf{y} = \sum\limits_{k=1}^{n} c_k \varphi_k - \mathbf{x} \neq 0$. Умножим его скалярно на любой вектор φ_i из нашего ортогонального базиса. Учитывая (2.97), получим

$$(\mathbf{y}, \ \varphi_i) = \left(\sum_{k=1}^{n} c_k \varphi_k, \ \varphi_i \right) - (\mathbf{x}, \ \varphi_i) = \| \varphi_i \|^2 (c_i - c_i) = 0,$$

и так как по предположению $\mathbf{y}$ — ненулевой вектор, то он ортогонален всем векторам базиса, что противоречит полноте базиса. Это противоречие и дока-

зывает (2.98). Аналогичное доказательство можно провести и для бесконечномерного базиса, но потребуются более тонкие аргументы для доказательства сходимости ряда.

В случае ортонормированного базиса $\|\varphi_k\| = 1$. Тогда, умножив скалярно обе части равенства (2.98) на вектор $\mathbf{x}$, найдем

$$\left(\sum_k c_k \varphi_k,\ \mathbf{x}\right) = \sum_k c_k (\varphi_k,\ \mathbf{x}) = \sum_k c_k (\mathbf{x},\ \varphi_k)^* = \sum_k c_k c_k^* =$$

$$= \sum_k |c_k|^2 = (\mathbf{x},\ \mathbf{x}) = \|\mathbf{x}\|^2.$$

Равенство

$$\sum_k |c_k|^2 = \|\mathbf{x}\|^2 \tag{2.100}$$

называется *равенством Парсеваля.* Оно выполняется для любого вектора только при полном ортонормированном базисе.

Два векторных пространства $\{\mathbf{x}\}$ и $\{\mathbf{y}\}$, называются изоморфными, если между элементами $\mathbf{x}$ и $\mathbf{y}$ можно установить взаимно однозначное соответствие и при этом выполняются следующие условия: если элементам $\mathbf{x}_k$ и $\mathbf{x}_l$ соответствуют элементы $\mathbf{y}_k$ и $\mathbf{y}_l$, то элементу $\mathbf{x}_k + \mathbf{x}_l$ соответствует $\mathbf{y}_k + \mathbf{y}_l$, элементу $\alpha\mathbf{x}_k$ соответствует $\alpha\mathbf{y}_k$ и $(\mathbf{x}_k,\ \mathbf{x}_l) = (\mathbf{y}_k,\ \mathbf{y}_l)$. Очевидно, что если пространство $\mathscr{A}$ изоморфно $\mathscr{B}$, а $\mathscr{B}$ изоморфно $\mathscr{C}$, то и $\mathscr{A}$ изоморфно $\mathscr{C}$.

В дальнейшем будет идти речь о различных функциональных пространствах. Одним из них является евклидово n-мерное пространство R_n. Его образует, например, множество периодических функций $X(t)$ с периодом T, представляемых в виде конечной тригонометрической суммы (разложения по ортогональному базису $\cos k(2\pi t/T)$, $\sin k(2\pi t/T)$:

$$x(t) = \sum_{k=k_1}^{k_2} \left(a_k \cos k\frac{2\pi t}{T} + b_k \sin k\frac{2\pi t}{T}\right). \tag{2.101}$$

Скалярное произведение в этом пространстве определяется следующей формулой:

$$(\mathbf{x}^{(1)},\ \mathbf{x}^{(2)}) = \frac{1}{T}\int_0^T x^{(1)}(t)\, x^{(2)}(t)\, dt = \frac{1}{2}\sum_{k=k_1}^{k_2} (a_k^{(1)} a_k^{(2)} + b_k^{(1)} b_k^{(2)}). \tag{2.102}$$

Такое же пространство образуют финитные функции, заданные на интервале $(0,\ T)$ или $(-T/2,\ T/2)$ и т. д., если на этом интервале их можно представить суммой (2.101).

Эти пространства изоморфны пространству n-мерных векторов, представляющих упорядоченные последовательности из n действительных чисел $\mathbf{x}(l_1,\ l_2,\ \ldots,\ l_n)$ со скалярным произведением

$$(\mathbf{x}^{(1)},\ \mathbf{x}^{(2)}) = \sum_{k=1}^{n} l_k^{(1)} l_k^{(2)}.$$

Важную роль в теории играет бесконечномерное пространство $L_2(T)$ финитных функций с интегрируемым квадратом или периодических функций с периодом T. Их можно представить на интервале $(0, T)$ рядом Фурье (2.101) при $k_2=\infty$. Скалярное произведение определяется также формулой (2.102). Это пространство изоморфно пространству l_2 бесконечных упорядоченных последовательностей чисел $\mathbf{x}=(l_1, l_2, \ldots)$, удовлетворяющих условию

$$\sum_{k=1}^{\infty} l_k^2 < \infty.$$

Функции с интегрируемым квадратом, заданные на бесконечной оси времени, образуют пространство $L_2(\infty)$, в котором скалярное произведение определяется выражением

$$(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \int_{-\infty}^{\infty} x^{(1)}(t)\, x^{(2)}(t)\, dt. \qquad (2.102а)$$

В функциональном анализе доказывается, что действительные пространства l_2, $L_2(T)$ и $L_2(\infty)$ изоморфны.

Аналогичные функциональные пространства образуют и комплекснозначные функции при соответствующем определении скалярного произведения. Так, для комплексного пространства $L_2(T)$

$$(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{1}{T}\int_0^T x^{(1)}(t)\cdot x^{(2)*}(t)\, dt. \qquad (2.102б)$$

Оно изоморфно пространству l_2 бесконечных последовательностей комплексных чисел $\mathbf{x}=(l_1, l_2, \ldots)$, удовлетворяющих условию $\sum_{k=1}^{\infty} |l_k|^2 < \infty$, для которого скалярное произведение определяется выражением

$$(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sum_{k=1}^{\infty} l_k^{(1)}\, l_k^{(2)*}.$$

Подпространство комплексного пространства $L_2(T)_{\text{анал}}$, содержащее только аналитические периодические сигналы $x(t)+ +\mathrm{i}\,\tilde{x}(t)$ (т. е. такие функции, в которых мнимая часть является преобразованием Гильберта от действительной части), также изоморфно всему комплексному пространству $L_2(T)$ и $l_2(\infty)$.

Заметим, что неравенство Коши—Буняковского—Шварца (2.93) в пространстве L_2 принимает вид

$$\left|\int x^{(1)}(t)\, x^{(2)*}(t)\right| \leqslant \sqrt{\int |x^{(1)}(t)|^2\, dt \int |x^{(2)}(t)|^2\, dt}, \qquad (2.103)$$

причем пределы интегрирования охватывают область определения пространства, например, от $-T/2$ до $T/2$ для пространства $L_2(T)$ или от $-\infty$ до $+\infty$ для $L_2(\infty)$.

Не всегда взаимно однозначное соответствие элементов двух пространств означает изоморфизм. Пусть $\mathscr{A}$ — действительное пространство $L_2(T)$, а $\mathscr{B}$ — пространство $L_2(T)_{\text{анал}}$ при том же периоде T. Между ними существует взаимно однозначное соответствие — комплексной функции из $\mathscr{B}$ соответствует в $\mathscr{A}$ ее действительная часть. Однако это соответствие не является изоморфизмом. Это видно из того, что двум ортогональным сигналам из пространства $\mathscr{A}$ [например, $\cos(2\pi lt/T)$ и $\sin(2\pi lt/T)$] могут соответствовать два неортогональных сигнала из пространства $\mathscr{B}$ (в этом примере

$$\cos\frac{2\pi lt}{T}+\mathrm{i}\sin\frac{2\pi lt}{T}=\mathrm{e}^{\mathrm{i}\frac{2\pi lt}{T}} \quad \text{и} \quad \sin\frac{2\pi lt}{T}-\mathrm{i}\cos\frac{2\pi lt}{T}=\mathrm{e}^{\mathrm{i}\left(\frac{2\pi lt}{T}-\frac{\pi}{2}\right)}$$

не ортогональны, так как

$$\int_{-T/2}^{T/2}\mathrm{e}^{\mathrm{i}\frac{2\pi lt}{T}}\,\mathrm{e}^{-\mathrm{i}\left(\frac{2\pi lt}{T}-\frac{\pi}{2}\right)}\,dt=\int_{-T/2}^{T/2}\mathrm{e}^{\mathrm{i}\frac{\pi}{2}}\,dt=\mathrm{i}T\neq 0.$$

С другой стороны, существуют ортогональные сигналы в пространстве $\mathscr{A}$, которым и в пространстве $\mathscr{B}$ соответствуют ортогональные сигналы (например, сигналам $\cos\frac{2\pi lt}{T}$ и $\cos\frac{2\pi mt}{T}$ $m\neq l$, в пространстве $\mathscr{B}$ соответствуют также ортогональные сигналы $\mathrm{e}^{\mathrm{i}\frac{2\pi lt}{T}}$ и $\mathrm{e}^{\mathrm{i}\frac{2\pi mt}{T}}$).

В теории передачи сигналов два сигнала, $x(t)$ и $y(t)$, называются *ортогональными в усиленном смысле*, если соответствующие им аналитические сигналы $x(t)+\mathrm{i}\tilde{x}(t)$ и $y(t)+\mathrm{i}\tilde{y}(t)$ также ортогональны. Легко убедиться, что для этого все четыре действительных сигнала, $x(t)$, $\tilde{x}(t)$, $y(t)$ и $\tilde{y}(t)$, должны быть попарно ортогональны. Заметим, что пара сопряженных по Гильберту сигналов (например, $x(t)$ и $\tilde{x}(t)$) всегда ортогональна, поскольку преобразование Гильберта сводится к повороту фаз на 90°. Кроме того, из ортогональности $x(t)$ с $y(t)$ следует также ортогональность $\tilde{x}(t)$ с $\tilde{y}(t)$. Поэтому для проверки ортогональности в усиленном смысле достаточно убедиться в ортогональности двух пар — $x(t)$ с $y(t)$ и $x(t)$ с $\tilde{y}(t)$ (или $\tilde{x}(t)$ с $y(t)$).

Рассмотрим еще пространство 2_n, элементами которого являются последовательности n чисел $\mathbf{a}(a_1, a_2, \ldots, a_n)$, где a_k принимают только значения 0 или 1. Сложение в этом пространстве производится поразрядно по модулю 2:

$$\mathbf{a}+\mathbf{b}=(a_1\oplus b_1,\ a_2\oplus b_2,\ \cdots,\ a_n\oplus b_n),$$

где знак $\oplus$ означает сложение по правилу (2.88). В отличие от остальных приведенных примеров, это пространство двоичных n-мерных векторов содержит лишь конечное число элементов. Так как каждый член последовательности может принимать только два значения, то всего различных последовательностей длины n существует 2^n.

Скалярное произведение в этом пространстве удобно задать формулой

$$(\mathbf{a},\ \mathbf{b}) = \left[\sum a_k b_k\right]^2,$$

где Σ — сумма в обычном смысле. Отсюда норма двоичного вектора.

$$\| \mathbf{x} \| = \sum_{k=1}^{n} a_k^2 = \sum_{k=1}^{n} a_k. \tag{2.104}$$

Последнее равенство записано на основании того, что $0^2=0$, $1^2=1$.

Таким образом, нормой двоичного вектора является количество содержащихся в нем единиц. Эту норму называют также *весом* вектора. Все свойства нормы при этом выполняются; в частности, норма равна нулю только для нулевого вектора (состоящего из одних нулей).

Ортонормированный базис в этом пространстве содержит n векторов:

$$(1,\ 0,\ 0,\ \cdots,\ 0);\ (0,\ 1,\ 0,\ 0,\ \cdots,\ 0)\ \cdots\ (0,\ 0,\ 0,\ \cdots,\ 0,\ 1).$$

Других ортонормированных базисов здесь нет, так как все остальные векторы (кроме нулевого) имеют норму не менее 2.

Расстояние между двумя векторами $d(\mathbf{a};\ \mathbf{b})$ по определению равно норме их разности:

$$d(\mathbf{a};\ \mathbf{b}) = \| \mathbf{a}-\mathbf{b} \| = \sum_{k=1}^{n} (a_k \ominus b_k) = \sum_{k=1}^{n} (a_k \oplus b_k), \tag{2.105}$$

где $\oplus$ и $\ominus$ означают сложение и вычитание по модулю 2, а последнее равенство написано на основании того, что сложение и вычитание по модулю 2 совпадают. Значение $a_k \oplus b_k$ равно 1 в том случае, когда $a_k \neq b_k$, если же $a_k = b_k$, то $a_k \oplus b_k = 0$. Таким образом, расстояние между двоичными векторами равно числу составляющих, в которых они различаются. Такое определение расстояния было введено Р. Хэммингом и называется *расстоянием Хэмминга*. Об этом пространстве речь пойдет в гл. 5.

Следует упомянуть еще об одном часто используемом функциональном пространстве — пространстве сигналов с конечной мощностью, т. е. таких, для которых

$$\lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2\, dt < \infty. \tag{2.106}$$

При этом энергия сигнала может быть бесконечной.

Множество таких сигналов безусловно образует линейное пространство. Его обозначают S или L_{2P}. Однако в нем не удается определить надлежащим образом скалярное произведение и норму. Если, как это часто делают, определить скалярное произведение формулой

$$(\mathbf{x}^{(1)},\ \mathbf{x}^{(2)}) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^{(1)}(t)\, x^{(2)}(t)\, dt \tag{2.107}$$

и, следовательно, норму

$$\| x \| = \left[\lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} | x(t) |^2 \, dt \right]^{1/2}, \qquad (2.108)$$

то норма ненулевых сигналов, имеющих конечную энергию, окажется равной нулю, что противоречит условию (2.91в). Поэтому не все полученные выше результаты применимы к пространству S. В частности, в нем не существует счетного ортогонального базиса.

Множество реализаций случайного процесса нередко образует линейное пространство. При этом скалярные произведения, нормы, расстояния оказываются случайными величинами. Заметим, что ортогональный базис состоит из детерминированных функций. В частности, в формуле (2.98) при разложении случайного процесса по ортогональному базису коэффициенты c_k — случайные величины, а φ_k — не случайные функции.

Случайные коэффициенты c_k в (2.98), вообще говоря, коррелированы между собой. Для решения многих вопросов полезно иметь разложение процесса по такому ортогональному базису, в котором коэффициенты c_k между собой не коррелированы. Разложение, удовлетворяющее этому условию, называют *каноническим*. Доказано, что для стационарных случайных процессов всегда существует хотя бы один базис (разный для различных процессов и определяемый корреляционной функцией процесса), при котором разложение оказывается каноническим.

2.7. ТЕОРЕМА КОТЕЛЬНИКОВА

Важное значение в теории передачи сигналов имеет пространство *сигналов с финитным спектром*, т. е таких сигналов $x(t)$ с интегрируемым квадратом, для которых преобразование Фурье $S_x(\mathrm{i}\,2\pi f) \equiv 0$ при $|f| > F$. Оно является подпространством пространства $L_2(\infty)$. Назовем его пространством $\mathcal{A}$, а пространство соответствующих спектров (комплексных функций S_x от аргумента f) назовем $\mathcal{B}$. Оно является пространством финитных функций, заданных на интервале $(-F, +F)$, и отличается от комплексного пространства $L_2(T)$ только обозначением аргумента и области определения.

Прямое и обратное преобразования Фурье определяют взаимно однозначное соответствие между пространствами $\mathcal{A}$ и $\mathcal{B}$, которое в данном случае является изоморфизмом. Действительно, из теории преобразования Фурье известно, что

$$\int_{-\infty}^{\infty} x(t)\, y(t)\, dt = \int_{-\infty}^{\infty} S_x(\mathrm{i}\,2\pi f)\, S_y^*(\mathrm{i}\,2\pi f)\, df, \qquad (2.109)$$

т. е. скалярные произведения соответствующих элементов пространств $\mathcal{A}$ и $\mathcal{B}$ одинаковы. Остальные условия изоморфности (см. с. 62) вытекают из линейности преобразования Фурье.

В пространстве $\mathcal{B}$ существует ортонормированный базис

$$\psi_k(f) = \frac{1}{\sqrt{2F}} \exp\left(-i\pi k \frac{f}{F}\right), \quad k = \cdots, -2, -1, 0, 1, 2, \cdots \tag{2.110}$$

Вследствие изоморфизма функциям $\psi_k(f)$ соответствуют некоторые функции $\hat{\varphi}_k(t)$ (обратные преобразования Фурье от $\psi_k(f)$), образующие в пространстве $\mathcal{A}$ также полный ортонормированный базис.

Найдем эти функции:

$$\hat{\varphi}_k(t) = \frac{1}{\sqrt{2E}} \int_{-F}^{F} \exp\left(-i\frac{\pi kf}{F}\right) \exp(i2\pi ft)\, df = \frac{1}{\sqrt{2F}} \times$$

$$\times \int_{-F}^{F} \exp\left[i2\pi\left(t - \frac{k}{2F}\right) f\right] df = \frac{1}{\sqrt{2F}} \left\{ \int_{-F}^{F} \cos\left[2\pi f (t - k\Delta t)\right] df + \right.$$

$$\left. + i \int_{-F}^{F} \sin\left[2\pi f (t - k\Delta t)\right] df \right\}, \tag{2.111}$$

где введено обозначение

$$\Delta t = \frac{1}{2F}. \tag{2.112}$$

Второй интеграл в (2.111) равен нулю. Первый же интеграл легко вычисляется и дает

$$\hat{\varphi}_k(t) = \frac{\sin 2\pi F (t - k\Delta t)}{\sqrt{2F}\,\pi\,(t - k\Delta t)}.$$

Ортогональность и полнота системы функций $\hat{\varphi}_k(t)$ не нарушатся, если каждую из них разделить на постоянную $\sqrt{2F}$, в результате чего получим ортогональный (не нормированный) базис в пространстве функций с финитным спектром:

$$\varphi_k(t) = \frac{\hat{\varphi}_k(t)}{\sqrt{2F}} = \frac{\sin 2\pi F (t - k\Delta t)}{2\pi F (t - k\Delta t)}, \tag{2.113}$$

где $\Delta t = 1/2F$; $k = \ldots, -2, -1, 0, 1, 2, \ldots$

Функции $\varphi_k(t)$ называются *функциями отсчета.* Три из них изображены на рис. 2.14. Они отличаются друг от друга только сдвигами по времени на интервалы, кратные Δt. Заметим, что при $t = n\Delta t$, где n — любое целое положительное или отрицательное число, все функции отсчета, кроме φ_n, равны нулю, а $\varphi_n = 1$.

На основании (2.98) любую функцию $x(t)$ с финитным спектром можно представить рядом

$$x(t) = \sum_{k=-\infty}^{\infty} c_k \frac{\sin 2\pi F (t - k\Delta t)}{2\pi F (t - k\Delta t)}, \tag{2.114}$$

называемым *рядом Котельникова.*

Для нахождения коэффициентом разложения c_k можно, конечно, воспользоваться формулой (2.97). Однако в данном случае существует более простой путь. Для определения коэффициента c_n положим $t=n\Delta t$. Тогда все члены ряда, кроме $c_n\varphi_n$, обратятся в нуль, а φ_n — в единицу, откуда

$$c_n = x(n\Delta t). \tag{2.115}$$

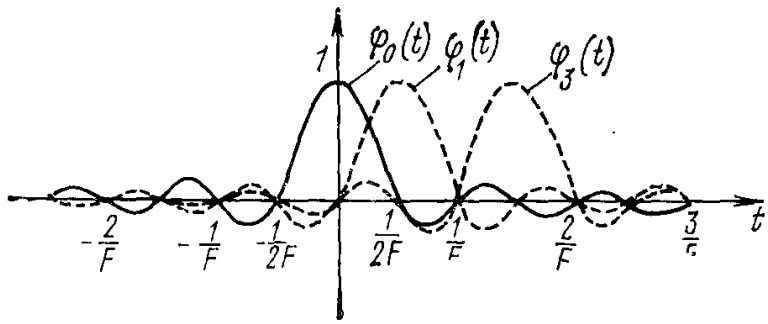


Рис. 2.14. Функция отсчета

Таким образом, коэффициенты ряда Котельникова представляют собой значения (отсчеты) функции $x(t)$, взятые в моменты времени, кратные Δt. Именно это обстоятельство определило исключительную важность ряда (2.132) в теории передачи сигналов и позволило В. А. Котельникову сформулировать и доказать следующую теорему:

Сигнал с финитным спектром можно точно восстановить (интерполировать) по его отсчетам, взятым через интервалы времени $\Delta t \leqslant 1/2F$, *где* F — *верхняя частота спектра сигнала.* Это осуществляется с помощью ряда

$$x(t) = \sum_{k=-\infty}^{\infty} x(k\Delta t)\frac{\sin 2\pi F_0(t-k\Delta t)}{2\pi F_0(t-k\Delta t)}, \tag{2.116}$$

где $F_0 = (1/2\Delta t) \geqslant F$. Фундаментальное значение теоремы Котельникова заключается в том, что она обосновывает возможность дискретизации по аргументу (времени) любых функций с финитным спектром. На ней основаны все методы импульсной модуляции. Эта теорема широко используется также в теоретических исследованиях.

Пусть для некоторых сигналов $x(t)$ с финитным спектром все отсчеты в точках $k\Delta t$, лежащих за пределами некоторого интервала времени длительностью T, равны нулю. Тогда ряд (2.116) вырождается в конечную сумму, число членов которой n равно числу отсчетных точек, умещающихся на интервале T:

$$n \approx T/\Delta t = 2FT. \tag{2.117}$$

Эта величина представляет, очевидно, размерность пространства. В теории связи ее называют *базой сигнала.*

Иногда полученный результат формулируют следующим образом: сигнал длительностью T, спектр которого не содержит частот выше F, полностью определяется заданием $2FT$ его отсчетов. Однако такая формулировка неточна и противоречива. Как известно из теории преобразования Фурье, спектр финитного сигнала не может быть финитным, так что сигналов, о которых говорится в приведенной фразе, в природе не существует. В частности, сигнал, представленный конечным числом членов ряда Котельникова, существует и за пределами интервала времени T, внутри которого находятся все ненулевые отсчеты. Это видно из того, что каждая функция отсчетов не финитна, хотя и затухает довольно быстро при удалении от своего максимума.

Тем не менее, на практике часто приходится иметь дело с финитными сигналами, энергия которых почти полностью сосредоточена внутри полосы частот $|f|\leqslant F$ и для таких сигналов нередко используют конечное число $2FT$ членов ряда Котельникова. Но в данном случае это представление является приближенным, и сумма такого конечного ряда отличается от функции $x(t)$ некоторой погрешностью $\varepsilon(t)$. Нетрудно показать, что относительный средний квадрат погрешности $\overline{\varepsilon^2(t)}/\overline{x^2(t)}$ (где волнистая черта означает усреднение по времени) равен доле энергии сигнала, лежащей за полосой частот F, и может быть очень малым, если $2FT\gg 1$.

Заметим, что такой финитный сигнал можно разложить и по тригонометрическому базису в обычный ряд Фурье (2.110) по частотам, кратным $1/T$. Если теперь отбросить малые члены, частоты которых превышают F, то легко убедиться, что число ненулевых членов такого ряда также примерно равно $2FT$. В этом ничего удивительного нет, так как оба эти представления можно трактовать как разложение функций в пространстве с размерностью $2FT$ по двум различным ортогональным базисам.

Дискретизированный по времени сигнал $x(t)$, т. е. последовательность его отсчетов, часто бывает представлен последовательностью очень коротких импульсов, площадь которых пропорциональна отсчетам $x(n\Delta t)$. Если пропустить такой импульс через идеальный фильтр нижних частот (ФНЧ) с полосой пропускания F, то на выходе восстановится сигнал $x(t)$. Действительно, импульсная реакция идеального ФНЧ равна $g(t)=\sin 2\pi Ft/(2\pi Ft)$, а последовательность входных импульсов (аппроксимируя их дельта-функциями) можно представить суммой $\sum\limits_{k=-\infty}^{\infty} x(k\Delta t)\,\delta(t-k\Delta t)$. Тогда на основании интеграла Дюамеля получим в момент t на выходе ФНЧ

$$\int\limits_{-\infty}^{\infty}\sum_{k=-\infty}^{\infty} x(k\Delta t)\,\delta(\tau-k\Delta t)\,\frac{\sin 2\pi F(\tau-t)}{2\pi F(\tau-t)}\,d\tau=$$

$$=\sum_{k=-\infty}^{\infty} x(k\Delta t)\,\frac{\sin 2\pi F(t-k\Delta t)}{2\pi F(t-k\Delta t)}=x(t). \qquad (2.118)$$

Реальный ФНЧ отличается от идеального, однако может к нему сколь угодно приблизиться, если допускается достаточная задержка. Неидеальность ФНЧ является дополнительным источником погрешности при восстановлении сигнала по импульсным отсчетам.

Для случайных процессов с финитным энергетическим спектром отсчеты (сечения) $x(k\Delta t)$ в (2.116) являются случайными числами. Последовательность $x(k\Delta t)$ полностью определяет реализацию сигнала $x(t)$. Если спектр сигнала в полосе $0<|f|<F$ равномерен, то отсчеты $x(k\Delta t)$, в соответствии с (2.48), некоррелированны. В этом случае разложение (2.116) является каноническим.

2.8. ИНФОРМАЦИОННЫЕ ПАРАМЕТРЫ СООБЩЕНИЙ И СИГНАЛОВ

В этом параграфе будут рассмотрены методы количественного определения информации, содержащейся в сообщении и передаваемой по каналу связи. Строгие методы количественного определения информации были предложены К. Шенноном в 1948 г. и привели к построению *теории информации*, являющейся математической основой теории связи, кибернетики и ряда смежных и даже довольно отдаленных отраслей науки.

Пусть некоторый источник дискретных сообщений посылает одно конкретное сообщение $a \in A$ из некоторого ансамбля A. Попытаемся найти определение количества информации, содержащейся в этом сообщении, исходя из следующих естественных требований:

1. Количество информации должно быть аддитивной мерой, т. е. количество информации в двух независимых сообщениях должно равняться сумме количеств информации в каждом из них.

2. Количество информации в сообщении о достоверном событии равно нулю.

3. Количество информации не должно зависеть от качественного содержания сообщения, в частности, от степени его важности для получателя, от возможных последствий его передачи, от эмоциональной окраски и т. д.

Первое требование настолько естественно, что в дополнительных обоснованиях не нуждается. Подчеркнем лишь, что речь идет о независимых сообщениях, когда прием одного из них никак не влияет на восприятие другого. Второе требование также легко понять, поскольку сообщение о достоверном событии не может ничего изменить в наших знаниях. Третье требование быть может не кажется столь очевидным. Однако оно обосновано необходимостью абстрагироваться от различных несущественных деталей, для того чтобы построить достаточно общую теорию. В качестве аналогии можно сослаться на то, что нельзя было бы построить научную механику, если по-разному определять массу тел, состоящих из различных веществ.

Итак, для определения количества информации в сообщении необходимо основываться только на таком параметре, который характеризует в самом общем виде сообщение a из ансамбля A. Таким параметром, очевидно, является вероятность $P(a)$ того, что источник посылает данное сообщение. Следовательно, количество информации $i(a)$, содержащееся в сообщении a, должно быть функцией от $P(a)$.

Дальнейшее уточнение искомого определения не составляет труда, если учесть первые два требования. Пусть a_1 и a_2 — два независимых сообщения. Вероятность того, что источник пошлет оба эти сообщения (одно за другим), равна $P(a_1, a_2) = P(a_1)P(a_2)$, а информация, содержащаяся в них, должна удовлетворять условию аддитивности, т. е. $i(a_1, a_2) = i(a_1) + i(a_2)$. Следовательно, необходимо найти функцию от вероятности P, обладающую тем свойством, что при перемножении двух аргументов значения функции складываются. Единственная такая функция —

это логарифмическая функция $i(a)=k\log P(a)$, где k — любая постоянная, а логарифм берется по любому основанию. Заметим, что при таком определении количества информации выполняется и второе требование — при $P(a)=1$ $i(a)=k\log 1=0$.

Что касается выбора коэффициента k и основания логарифма, то они определяются только удобством. Заметим, что изменение основания логарифма также можно свести к изменению коэффициента, поскольку $k_1\log_m P(a)=k_1\log_n P(a)\log_m n=k_2\log_n P(a)$, так что замена основания логарифма с n на m сводится к умножению коэффициента на $\log_m n$. Поэтому переход от одного основания к другому означает лишь изменение единицы количества информации.

Для того чтобы количество информации измерялось неотрицательным числом, будем всегда выбирать $k=-1$, поскольку $P(a)\leqslant 1$ и $\log P(a)\leqslant 0$ (если основание логарифма больше единицы). Поэтому

$$i(a)=-\log P(a)=\log\frac{1}{P(a)}. \qquad (2.119)$$

Основание логарифма в (2.119) чаще всего выбирают равным 2. Полученная при этом единица информации носит название *двоичная единица* или *бит*. Она равна количеству информации в сообщении о событии, происходящем с вероятностью 0,5, т. е. таком, которое может с равной вероятностью произойти или не произойти. Такая единица на практике наиболее удобна вследствие широкого использования двоичных кодов в вычислительной технике и связи. В теоретических исследованиях иногда применяют натуральный логарифм, измеряя информацию в *натуральных единицах*. Натуральная единица в $\log_2 e\approx 1{,}443$ раза больше двоичной. Мы будем пользоваться в основном двоичными единицами, и в дальнейшем обозначение log будет означать двоичной логарифм.

Итак, количество информации в сообщении тем больше, чем оно менее вероятно или, другими словами, чем оно более неожиданно.

Если источник передает последовательность зависимых между собой сообщений, то получение предшествующих сообщений может изменить вероятность последующего, а следовательно, и количество информации в нем. Оно должно определяться по условной вероятности передачи данного сообщения a_n при известных предшествовавших a_{n-1}, a_{n-2}, ...,:

$$i(a_n\mid a_{n-1}, a_{n-2}, \cdots)=\log\frac{1}{P(a_n\mid a_{n-1}, a_{n-2}\cdots)}. \qquad (2.120)$$

Определенное выше количество информации является случайной величиной, поскольку сами сообщения случайные. Его распределение вероятностей определяется распределением вероятностей сообщений в данном ансамбле. Для характеристики же всего ансамбля (или источника) сообщений используется математи-

ческое ожидание количества информации, называемое *энтропией* и обозначаемое $H(A)$:

$$H(A) = M\left\{\log \frac{1}{P(a)}\right\}. \qquad (2.121)$$

Здесь математическое ожидание, как всегда, обозначает усреднение по всему ансамблю сообщений. При этом должны учитываться все вероятностные связи между различными сообщениями.

Чем больше энтропия источника, тем больше степень неожиданности передаваемых им сообщений в среднем, т. е. тем более неопределенным является ожидаемое сообщение. Поэтому энтропию часто называют мерой неопределенности сообщений. При этом имеется в виду неопределенность, существующая до того, как сообщение передано. После приема сообщения (если оно заведомо принимается верно) всякая неопределенность устраняется. Это позволяет трактовать количество информации как меру уменьшения неопределенности. Можно характеризовать энтропию также как меру разнообразия выдаваемых источником сообщений.

Энтропия является основной характеристикой источника. Чем она выше, тем труднее запомнить (записать) сообщение или передать его по каналу связи. Так, например, далее будет показано, что во многих случаях необходимая затрата энергии на передачу сообщения пропорциональна его энтропии.

Перечислим основные свойства энтропии.

1. Энтропия неотрицательна, она равна нулю только для «вырожденного» ансамбля, когда одно сообщение передается с вероятностью 1, а остальные имеют нулевую вероятность.

2. Энтропия аддитивна. Это значит, в частности, что если рассматривать последовательность из n сообщений как одно «укрупненное» сообщение, то энтропия источника таких укрупненных сообщений будет в n раз больше энтропии исходного источника.

3. Если ансамбль содержит K различных сообщений, то $H(A) \leqslant \log K$, причем равенство имеет место только тогда, когда все сообщения передаются равновероятно и независимо. Число K называется *объемом алфавита* источника.

Свойство 1 вытекает из выражения (2.121), если учесть, что $\frac{1}{P(a)} \geqslant 1$ и, следовательно, $\log \frac{1}{P(a)} \geqslant 0$, а математическое ожидание неотрицательной величины не может быть отрицательным. Свойство 2 легко понять исходя из того, что аддитивность была положена в основу определения энтропии. Это соображение, конечно, не является доказательством, однако свойство 2 может быть показано совершенно строго.

Приведем доказательство свойства 3 для источника независимых сообщений.

Пусть ансамбль содержит K различных сообщений. Если сообщения передаются статистически независимо друг от друга [1], с различными вероятностями,

[1] Такой источник называется источником без памяти.

то формула (2.121) принимает простой вид:

$$H(A) = \sum_{k=1}^{K} P(a_k) \log \frac{1}{P(a_k)}. \qquad (2.122)$$

Рассмотрим разность:

$$H(A) - \log K = \sum_{k=1}^{K} P(a_k) \log \frac{1}{P(a_k)} - \log K =$$

$$= -\sum_{k=1}^{K} P(a_k) \left[\log \frac{1}{P(a_k)} - \log K\right] = \sum_{k=1}^{K} P(a_k) \log \frac{1}{KP(a_k)} =$$

$$= \log e \sum_{k=1}^{K} P(a_k) \ln \frac{1}{KP(a_k)}. \qquad (2.123)$$

Воспользуемси известным неравенством [1]

$$\ln x \leqslant x - 1, \qquad (2.124)$$

справедливым для любого положительного x. Тогда

$$H(A) - \log K \leqslant (\log e) \sum_{k=1}^{K} P(a_k) \left[1 - \frac{1}{KP(a_k)}\right] = (\log e)[1 - 1] = 0,$$

причем равенство имеет место только при $KP(a_k) = 1$, что и требовалось доказать.

В частности, для *двоичного* источника без памяти, когда $K=2$, энтропия максимальна при $P(a_1) = P(a_2) = 0{,}5$ и равна $\log 2 = 1$ бит. Зависимость энтропии этого источника от $P(a_1) = 1 - P(a_2)$ показана на рис. 2.15.

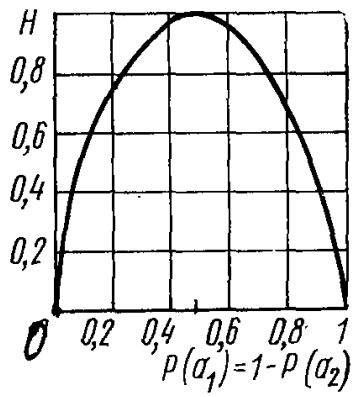


Рис. 2.15. Энтропии двоичного источника без памяти

Для источника, сообщения которого образуют простую цепь Маркова, вероятность каждого сообщения a_k целиком определена, если известно переданное непосредственно перед ним сообщение a_i, и формула (2.121) принимает следующий вид:

$$H(A) = \sum_{k=1}^{K} \sum_{i=1}^{K} P(a_i) P(a_k | a_i) \log \frac{1}{P(a_k | a_i)}, \qquad (2.125)$$

где $P(a_k|a_i)$ обозначает условную вероятность передачи a_k, если предыдущим сообщением было a_i, а $P(a_i)$ — безусловную (средиюю) вероятность передачи

[1] Для доказательства этого неравенства достаточно проверить, что функция $\ln x + 1 - x$ имеет отрицательную вторую пронзводную, а первая пронзводная равна нулю при $x=1$, когда неравенство переходит в равенство.

a_i. Заметим, что $\sum_{k=1}^{K} P(a_i)P(a_k|a_i)=P(a_k)$ — безусловная вероятность сообщения a_k.

Аналогично можно представить энтропию источника зависимых сообщений, если зависимость простирается не только на предыдущее сообщение. В общем случае при любых зависимостях между сообщениями энтропию можно определить из следующих соображений. Рассмотрим последовательность из n сообщений, которую обозначим $a_k^{[n]}$. Пусть ее вероятность равна $P(a_k^{[n]})$. Тогда содержащаяся в ней информация, согласно (2.119), равна $\log \frac{1}{P\left(a_k^{[n]}\right)}$, так что на каждое сообщение приходится $\frac{1}{n}\log\frac{1}{P\left(a_k^{[n]}\right)}$, а усреднение по всем таким последовательностям даст $\frac{1}{n}\sum_{k=1}^{K^n} P\left(a_k^{[n]}\right)\log\frac{1}{P\left(a_k^{[n]}\right)}$. Для устранения влияния предыдущих последовательностей устремим n к бесконечности. Тогда

$$H(A)=\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{K^n}P\left(a_k^{[n]}\right)\log\frac{1}{P\left(a_k^{[n]}\right)}. \qquad (2.126)$$

В теории информации доказывается, что энтропия источника зависимых сообщений всегда меньше энтропии источника независимых сообщений при том же объеме алфавита и тех же безусловных вероятностях сообщений.

Пусть, например, источник выдает последовательность букв из алфавита объемом $K=32$. Если буквы выбираются равновероятно и независимо друг от друга, то энтропия источника $\log K=5$ бит. Таким источником могла бы быть обезьяна, нажимающая в хаотическом порядке клавиши пишущей машинки.

Если буквы передаются не хаотически, а составляют связный русский текст, то они оказываются неравновероятными (например, буква *А* передается значительно чаще, чем *Ф*) и, главное, зависимыми (так, после гласных не может появиться *Ь*; мала вероятность сочетания более трех согласных подряд; вероятность последовательности, не образующей осмысленных слов, практически равна нулю). Если рассматривать ансамбль текстов русской художественной прозы, то энтропия оказывается менее 1,5 бит на букву. Еще меньше, около 1 бит на букву, энтропия ансамбля поэтических произведений, так как в них имеются дополнительные вероятностные связи, обусловленные ритмом и рифмами. Если же рассматривать в качестве источника поток телеграмм, то его энтропия обычно не превышает 0,8 бит на букву, поскольку тексты довольно однообразны.

Величина

$$\varkappa=(\log K-H(A))/\log K \qquad (2.127)$$

называется избыточностью источника с объемом алфавита K. Она показывает, какая доля максимально возможной при этом алфавите энтропии не используется источником.

Отметим некоторые свойства длинных последовательностей сообщений стационарного источника. Общее число N последовательностей длины n равно K^n. Будем для сокращения называть их n-последовательностями. Для источника равновероятных и не-

зависимых сообщений все n-последовательности равновероятны, т. е. вероятность каждой из них $P_n = K^{-n}$, откуда $\frac{1}{n}\log\frac{1}{P_n} = \log K = H_{\text{макс}}(A)$.

Если сообщения источника A неравновероятны и (или) зависимы, то, очевидно, среди n-последовательностей будут более вероятные и менее вероятные. Предположим, что составлен список всех n-последовательностей, расположенных в порядке убывания их вероятностей. Зададимся некоторым малым положительным числом δ и будем отсчитывать n-последовательности в порядке убывания их вероятностей, пока суммарная вероятность не превысит $1-\delta$. Назовем отсчитанные последовательности типичными и обозначим их число $N_{\text{тип}}$. Остальные нетипичные n-последовательности имеют суммарную вероятность меньше δ.

Для широкого класса стационарных источников в теории информации доказана следующая теорема об асимптотической равновероятности:

При любых $\delta > 0$ *и* $\varepsilon > 0$ *существует такое* n_0, *что при* $n > n_0$ *все* n-*последовательности можно разделить на типичные и нетипичные, причем суммарная вероятность нетипичных* n-*последовательностей меньше* δ, *а для вероятности* P_i *любой типичной* n-*последовательности справедливо неравенство*

$$\left|\frac{1}{n}\log\frac{1}{P_i} - H(A)\right| < \varepsilon. \qquad (2.128)$$

Отсюда следует, что при $n > n_0$ число типичных n-последовательностей $N_{\text{тип}} \approx 2^{nH(A)}$.

Эта теорема, являющаяся, по существу, выражением закона больших чисел, широко используется в теории информации. Она позволяет рассматривать любой расширенный источник $A^{[n]}$ при достаточно большом n как источник почти независимых и почти равновероятных типичных n-последовательностей с объемом алфавита $N_{\text{тип}} = 2^{nH(A)}$, в которые лишь очень редко, с вероятностью δ, вклинивается какая-нибудь нетипичная n-последовательность, причем δ стремится к нулю с увеличением n.

Интересно отметить, что при $H(A) < \log K$ и достаточно большом n типичные n-последовательности составляют лишь ничтожную долю всех возможных K^n последовательностей. Действительно, $\frac{N_{\text{тип}}}{K^n} \approx \frac{2^{nH(A)}}{K^n} = 2^{n[H(A)-\log K]}$, и если $H(A) < \log K$, то это отношение с увеличением n экспоненциально стремится к нулю.

Некоторые источники передают сообщения с фиксированной скоростью, затрачивая в среднем время T на каждое сообщение. *Производительностью* (в бит на секунду) такого источника $H'(A)$ назовем суммарную энтропию сообщений, переданных за единицу времени:

$$H'(A) = \frac{1}{T}H(A). \qquad (2.129)$$

У других источников скорость передачи сообщений определяется самой системой связи. Для таких источников с *управляемой скоростью* производительность может регулироваться в широких пределах, путем изменения величины T. Примером источника с фиксированной скоростью является датчик на космическом корабле, передающий каждые 10 с значения температуры, давления воздуха и т. д. Источником с управляемой скоростью является написанный на бумаге текст, подлежащий передаче по телеграфу.

2.9. ВЗАИМНАЯ ИНФОРМАЦИЯ

Энтропия ансамбля характеризует среднее количество полной информации, содержащейся в сообщении. Определим теперь информацию, содержащуюся в одном ансамбле относительно другого, например в принятом сигнале относительно переданного сообщения. Для этого рассмотрим объединение двух дискретных ансамблей, A и B, вообще говоря, зависимых. Его можно интерпретировать как пару ансамблей сообщений, либо как ансамбли сообщения и сигнала, с помощью которого сообщение передается, либо как ансамбли сигналов на входе и выходе канала связи и т. д. Пусть $P(a_k, b_l)$ — совместная вероятность реализаций a_k и b_l. Совместной энтропией ансамблей A и B будем называть

$$H(A, B) = M\left\{\log \frac{1}{P(a_k, b_l)}\right\}. \tag{2.130}$$

Введем также понятие условной энтропии

$$H(A \mid B) = M\left\{\log \frac{1}{P(a_k \mid b_l)}\right\}, \tag{2.131}$$

где $P(a_k|b_l)$ — условная вероятность a_k, если имеет место b_l; здесь математические ожидания берутся по объединенному ансамблю AB. В частности, для источников без памяти

$$H(A \mid B) = \sum_k \sum_l P(a_k, b_l) \log \frac{1}{P(a_k \mid b_l)}. \tag{2.132}$$

Из теоремы умножения вероятностей $P(a, b) = P(a)P(b|a) = P(b)P(a|b)$ следует, что

$$H(A, B) = H(A) + H(B \mid A) = H(B) + H(A \mid B). \tag{2.133}$$

Для условной энтропии справедливо двойное неравенство

$$0 \leqslant H(A \mid B) \leqslant H(A). \tag{2.134}$$

При этом равенство $H(A|B)=0$, как видно из (2.131), имеет место в том случае, когда при каждом значении b_l условная вероятность одной реализации $P(a_i|b_l)=1$, а для всех остальных реализаций $P(a_k|b_l)=0$. Это означает, что, зная реализацию B, можно точно установить реализацию A. Другими словами, B содержит полную информацию об A.

Другой крайний случай, когда $H(A|B)=H(A)$, имеет место, если $P(a_k|b_l)=P(a_k)$ при всех a и b. Последнее равенство означает, что события A и B независимы. В этом случае знание реализации B не уменьшает неопределенности A, т. е. B не содержит никакой информации об A.

В общем случае условная энтропия $H(A|B)$ меньше безусловной $H(A)$ и знание реализации B снижает в среднем первоначальную неопределенность A. Естественно назвать разность $H(A)-H(A|B)$ *количеством информации, содержащейся в B относительно A*. Ее называют также *взаимной информацией* между A и B и обозначают $I(A, B)$:

$$I(A, B)=H(A)-H(A \mid B). \tag{2.135}$$

Подставляя сюда (2.121) и (2.131), выразим взаимную информацию через распределения вероятностей:

$$I(A, B)=M\left\{\log\frac{1}{P(a_k)}\right\}-M\left\{\log\frac{1}{P(a_k \mid b_l)}\right\}=M\left\{\log\frac{P(a_k \mid b_l)}{P(a_k)}\right\}. \tag{2.136}$$

Если воспользоваться теоремой умножения $P(a_k, b_l)=P(b_l)P(a_k|b_l)$, то можно записать $I(A, B)$ в симметричной форме:

$$I(A, B)=M\left\{\log\frac{P(a_k, b_l)}{P(a_k)P(b_l)}\right\}. \tag{2.137}$$

Сформулируем основные свойства взаимной информации:

1. $I(A, B)\geqslant 0$, (2.138)

причем равенство имеет место тогда и только тогда, когда A и B независимы между собой. Это следует из определения (2.135) и неравенства (2.134).

2. $I(A, B)=I(B, A)$, (2.139)

т. е. B содержит столько же информации относительно A, сколько A содержит относительно B. Это свойство вытекает из симметрии выражения (2.137). Поэтому можно также записать

$$I(A, B)=H(B)-H(B \mid A). \tag{2.140}$$

3. $I(A, B)\leqslant H(A)$, (2.141)

причем равенство имеет место, когда по реализации B можно однозначно восстановить A. Это следует из (2.135) и (2.134).

4. $I(A, B)\leqslant H(B)$, (2.142)

причем равенство имеет место, когда по реализации A можно точно установить реализацию B. Это вытекает из (2.139) и (2.141).

5. Полагая в (2.135) $B=A$ и учитывая, что $H(A|A)=0$, получим

$$I(A, A)=H(A). \tag{2.143}$$

Это позволяет интерпретировать энтропию источника как его *собственную информацию,* т. е. информацию, содержащуюся в ансамбле A о самом себе.

Пусть, например, A — ансамбль дискретных сообщений, а B — ансамбль дискретных сигналов, в которые преобразуются сообщения A. Тогда $I(A, B)=H(A)$ в том и только в том случае, когда преобразование $A\to B$ обратимо. При необратимом преобразовании $I(A, B)<H(A)$ и разность $H(A)-I(A, B)=H(A|B)$ можно назвать *потерей информации* при преобразовании $A\to B$. Ее называют также *ненадежностью.* Таким образом, информация не теряется только при обратимых преобразованиях.

Если T — среднее время передачи одного сообщения, то, разделив формулы (2.130)—(2.142) на T и обозначая

$$H'(A \mid B)=\frac{1}{T}H(A \mid B),\quad I'(A, B)=\frac{1}{T}I(A, B)$$

и т. д., получим соответствующие равенства для энтропий и количества информации, рассчитанных не на одно сообщение, а на единицу времени. Величина $I'(A, B)$ называется *скоростью передачи информации* от A к B (или наоборот).

Если, например, U — ансамбль сигналов на входе дискретного канала, а Z — ансамбль сигналов на его выходе, то скорость передачи информации по каналу выразится следующим образом:

$$I'(U, Z)=H'(U)-H'(U \mid Z)=H'(Z)-H'(Z \mid U).$$

Эти соотношения наглядно иллюстрируются на рис. 2.16. Здесь $H'(U)$ — производительность источника передаваемого сигнала U, а $H'(Z)$ — «производительность» канала, т. е. полная собственная информация в принятом сигнале за единицу времени. Величина $H'(U|Z)$ представляет собой скорость «утечки» информации при прохождении через канал, а $H'(Z|U)$ — скорость передачи посторонней информации, не имеющей отношения к U и создаваемой присутствующими в канале помехами. Соотношение между $H'(U|Z)$ и $H'(Z|U)$ зависит от свойств канала. Так, например, при передаче телефонного сигнала по каналу с узкой полосой пропускания, недостаточной для удовлетворительного воспроизведения сигнала, и с низким уровнем помех теряется часть полезной информации, но почти не получается бесполезной. В этом случае $H'(U|Z)\gg H'(Z|U)$. Если же сигнал воспроизводится хорошо, но в паузах ясно прослушиваются «наводки» от соседнего телефонного канала, то, почти не теряя полезной информации, можно получить много дополнительной, как правило, бесполезной информации и $H'(U|Z)\ll H'(Z|U)$.

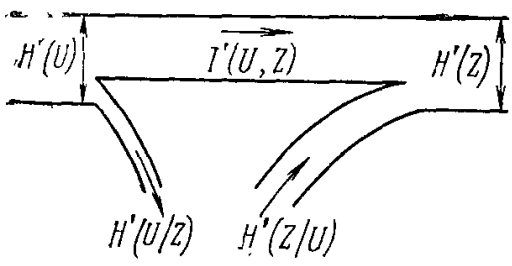


Рис. 2.16. Иллюстрация передачи информации по каналу с помехами

2.10. ЭФФЕКТИВНОЕ КОДИРОВАНИЕ ДИСКРЕТНЫХ СООБЩЕНИЙ

Применим полученные результаты к проблеме кодирования дискретных сообщений. Пусть A — источник последовательности элементарных сообщений (знаков) с объемом алфавита K и производительностью $H'(A)$. Для передачи по дискретному каналу нужно преобразовать сообщения в последовательность кодовых символов B так, чтобы эту кодовую последовательность можно было затем однозначно декодировать[1]. Для этого необходимо, чтобы скорость передачи информации от источника к коду $I'(A, B)$ равнялась производительности источника $H'(A)$. Но, с другой стороны, согласно (2.142) $I'(A, B) \leqslant H'(B)$. Следовательно, необходимым условием для кодирования является $H'(B) \geqslant H'(A)$ или, обозначая через $T_к$ длительность кодового символа, а через $T_с$ длительность элементарного сообщения, $\frac{1}{T_к} H(B) \geqslant \frac{1}{T_с} H(A)$, или

$$v_к H(B) \geqslant v_с H(A), \tag{2.144}$$

где $v_к = 1/T_к$ — число кодовых символов, а $v_с = 1/T_с$ — число сообщений, передаваемых в секунду. Будем рассматривать для простоты только двоичный код, при котором алфавит B состоит из символов 0 и 1. Тогда $H(B) \leqslant \log 2 = 1$ бит. Поэтому необходимое условие (2.144) сводится к тому, что

$$\frac{v_к}{v_с} \geqslant H(A). \tag{2.145}$$

Но $v_к/v_с$ представляет среднее число кодовых символов, приходящихся на одно элементарное сообщение. Таким образом, для возможности кодирования и однозначного декодирования сообщения необходимо, чтобы среднее число двоичных символов на сообщение было не меньше энтропии $H(A)$. Является ли это условие достаточным?

Одна из основных теорем теории информации утверждает, что оно «почти достаточно». Точнее, содержание *теоремы кодирования для источника* заключается в том, что, *передавая двоичные символы со скоростью* $v_к$ $\frac{\text{симв}}{\text{с}}$, *можно закодировать сообщения так, чтобы передавать их со скоростью*

$$v_с = \frac{v_к}{H(A)} - \varepsilon \text{ (сообщений в секунду)}, \tag{2.146}$$

где ε — *сколь угодно малая величина.*

Эта теорема почти тривиальна, если источник передает сообщения независимо и равновероятно. В этом случае $H(A) = \log K$, и если еще к тому же K — целая степень двух ($K = 2^n$), то $H(A) = n$. С другой стороны, используя для передачи каждого сигнала последовательность из n двоичных символов, мы получим $2^n = K$ различных последовательностей и можем каждому сигналу сопоставить одну из кодовых последовательностей, так что (2.146) выполняется даже при $\varepsilon = 0$.

Таким же образом можно закодировать сообщения любого источника с объемом алфавита K, затрачивая $v_0 = \log K$ двоичных символов на элементарное сообщение. Однако сформулированная теорема утверждает, что возможно более экономное кодирование, с затратой $v \approx H(A)$ символов на сообщение. Относительная экономия символов при этом окажется равной $(v_0 - v)/v_0 \approx$ [см. ф-лу (2.127)]. Таким образом, избыточность определяет достижимую степень «сжатия сообщения».

Доказательство этой теоремы можно найти в [20]. Здесь же ограничимся несколькими примерами.

Так, если элементарными сообщениями являются русские буквы ($K = 32 = 2^5$) и они передаются равновероятно и независимо, то $H(A) = 5$ бит. Каждую

---

[1] В этом параграфе предполагается, что последовательность кодовых символов принимается без ошибок. Поэтому рассматриваемую ниже теорему часто называют теоремой кодирования для канала без помех.

букву можно закодировать последовательностью из пяти двоичных символов, поскольку существует 32 такие последовательности. Разумеется, таким же равномерным кодом можно закодировать и буквы в связном русском тексте, и именно так часто поступают на практике. Но можно обойтись значительно меньшим числом символов на букву. Как указывалось выше, для русского литературного текста $H(A) \approx 1{,}5$ бит и, следовательно, возможен способ *эффективного кодирования* (или кодирования со *сжатием сообщения*), при котором в среднем на букву русского текста будет затрачено немногим более 1,5 двоичных символа, т. е. на 70% меньше, чем при примитивном коде.

Существует довольно много способов сжатия сообщений или сокращения избыточности текста. *Так, напр. эта фр. напис. сокращ., и тем не м. мож. надеят., что чит-ль пойм. ее прав.* В предыдущей фразе удалось уменьшить число букв (а следовательно, и символов, если ее кодировать равномерным кодом) почти на 40%.

Другая возможность, основанная на свойстве асимптотической равновероятности, заключается в том, чтобы кодировать не отдельные буквы, а целые слова. В достаточно большом словаре имеется около 10 000 слов, содержащих в среднем по 7 букв. Считая, что в среднем каждое словое может иметь 3 грамматические формы, нам придется закодировать около 30 000 типичных слов. Если применить равномерный код, то на каждое слово придется затратить 15 двоичных символов ($2^{15}=32\,768$). Таким образом, в среднем на букву придется немногим больше двух символов, т. е. сжатие достигнет почти 60%, вместо теоретических 70%. Слова, не вошедшие в словарь, придется передавать обычным способом, затрачивая пять символов на букву. Но поскольку такие нетипичные слова будут встречаться крайне редко, это не внесет заметной поправки в общий баланс.

Дальнейшее сжатие сообщений возможно путем применения неравномерного кода, если более короткие последовательности используются для более частых слов и более длинные — для более редких. Заметим, что эта идея неравномерного кодирования впервые нашла применение в телеграфном коде Морзе, в котором наиболее короткие комбинации использованы для часто встречающихся букв (е, и, т, с, а).

Разработано много методов эффективного кодирования для различных источников. Почти все они основаны на тех же двух принципах — укрупнения сообщений (аналогично переходу от букв к словам) и применения неравномерного кода. Заметим, что задача эффективного кодирования наиболее актуальна не для передачи текста, а для других источников со значительно большей избыточностью. К ним относятся, например, телевизионные передачи (промышленное телевидение) и некоторые телеметрические системы, в которых возможно сжатие в десятки раз.

2.11. ИНФОРМАЦИЯ В НЕПРЕРЫВНЫХ СИГНАЛАХ

Обобщим теперь понятия энтропии и взаимной информации на ансамбли непрерывных сигналов. Пусть S — случайная величина (сечение или отсчет случайного сигнала), определенная в некоторой непрерывной области, и ее распределение вероятностей характеризуется плотностью $w(s)$.

Разобьем область значений величины S на небольшие интервалы протяженностью Δs. Вероятность того, что $s_k < S < s_k + \Delta s$, приблизительно равна $w(s)\Delta s$, причем приближение тем точнее, чем меньше интервал Δs. Степень неожиданности такого события равна $\log \dfrac{1}{P\{s_k < S < s_k + \Delta s\}} \approx \log \dfrac{1}{w(s)\,\Delta s}$. Если не уточнять значение S в пределах конечного интервала Δs, а заменить их значе-

ниями s_k в начале интервала, то непрерывный ансамбль заменится дискретным, а его энтропия определится как

$$M\left\{\log\frac{1}{w(s)\,\Delta s}\right\}.$$

Будем теперь увеличивать точность определения значения s, уменьшая интервал Δs. В пределе, при $\Delta s\to 0$, должна получиться энтропия непрерывной случайной величины:

$$H(S)=\lim_{\Delta s\to 0} M\left\{\log\frac{1}{w(s)\,\Delta s}\right\}=\lim_{\Delta s\to 0} M\left\{\log\frac{1}{w(s)}+\log\frac{1}{\Delta s}\right\}=M\left\{\log\frac{1}{w(s)}\right\}+$$

$$+\lim_{\Delta s\to 0}\log\frac{1}{\Delta s}. \qquad (2.147)$$

Второй член в полученном выражении стремится к $+\infty$ и совершенно не зависит от распределения вероятностей S. Это значит, что собственная информация любой непрерывной случайной величины бесконечно велика. В этом нет ничего неожиданного, так как для того, чтобы точно задать значение S, например, в виде десятичной дроби, необходимо сообщить бесконечное количество цифр.

Тем не менее, как сейчас будет показано, взаимная информация между двумя непрерывными ансамблями, как правило, остается конечной. Такова будет, в частности, взаимная информация между переданным и принятым сигналами, так что по каналу связи информация передается с конечной скоростью[1].

Обратим внимание на первый член в (2.147). Он является конечным и определяется плотностью распределения вероятности $w(s)$. Его называют *дифференциальной энтропией* и обозначают $h(S)$:

$$h(S)=M\left\{\log\frac{1}{w(S)}\right\}=\int_S w(s)\log\frac{1}{w(s)}\,ds. \qquad (2.148)$$

Дифференциальную энтропию, в отличие от обычной энтропии дискретного ансамбля, нельзя рассматривать как меру собственной информации. Она не обладает многими свойствами обычной энтропии, в частности, может принимать и отрицательные значения. Информационный смысл имеет не сама дифференциальная энтропия, а разность двух дифференциальных энтропий, чем и объясняется ее название. Впрочем, свойство аддитивности сохраняется и для дифференциальной энтропии, т. е. дифференциальная энтропия нескольких сечений случайного процесса равна сумме их дифференциальных энтропий, вычисляемых, конечно, с учетом вероятностных зависимостей между сечениями.

Попытаемся теперь определить с помощью предельного перехода взаимную информацию между двумя непрерывными случайными величинами, S и U. Разбив области определения S и U соответственно на небольшие интервалы Δs и Δu, заменим эти непрерывные величины дискретными так же, как это делалось при выводе формулы (2.147).

---

[1] К вопросу об определении собственной информации (энтропии) непрерывного ансамбля сообщений мы вернемся в гл. 6.

Исходя из выражения (2.137), можно определить взаимную информацию между непрерывными величинами S и U:

$$I(S, U) = \lim_{\Delta s \to 0} M\left\{\log \frac{w(s, u)\,\Delta s\,\Delta u}{w(s)\,\Delta s w(u)\,\Delta u}\right\} = M\left\{\log \frac{w(s, u)}{w(s)\,w(u)}\right\}. \qquad (2.149)$$

При этом предельном переходе никаких явных бесконечностей не появилось, и действительно, в обычных случаях взаимная информация (2.149) оказывается конечной. С помощью простых преобразований ее можно представить и в таком виде:

$$I(S, U) = M\left\{\log \frac{w(u)\,w(s \mid u)}{w(s)\,w(u)}\right\} = M\left\{\log \frac{1}{w(s)} - \log \frac{1}{w(s \mid u)}\right\} =$$

$$= M\left\{\log \frac{1}{w(s)}\right\} - M\left\{\log \frac{1}{w(s \mid u)}\right\} = h(S) - h(S \mid U). \qquad (2.150)$$

Здесь $h(S)$ — определенная ранее дифференциальная энтропия S, а $h(S \mid U) = M\left\{\log \frac{1}{w(s \mid u)}\right\}$ *условная дифференциальная энтропия*. Формула (2.150) имеет ту же форму, что и (2.135), и отличается лишь заменой энтропии дифференциальной энтропией. Легко убедиться, что основные свойства взаимной информации (2.138) и (2.139) остаются справедливыми и в данном случае.

Аналогично по формуле (2.149) или (2.150) определяется и взаимная информация между двумя случайными n-мерными векторами, но фигурирующие в этой формуле плотности вероятностей оказываются многомерными.

Для нахождения взаимной информации между случайными процессами чаще всего используют их разложение по некоторому ортогональному базису. Взаимная информация между процессами $S(t)$ и $U(t)$ совпадает со взаимной информацией между случайными векторами $\mathbf{X}$ и $\mathbf{Y}$, составляющими которых являются коэффициенты разложения соответственно $S(t)$ и $U(t)$. Задача несколько облегчается, если применить не произвольное, а каноническое разложение (см. с. 66), так как при этом составляющие каждого вектора между собой не коррелированы.

В качестве примера, который понадобится в дальнейшем, найдем дифференциальную энтропию случайной величины X с нормальным распределением вероятности:

$$w(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-a)^2\right], \qquad (2.151)$$

где a — математическое ожидание, а σ^2 — дисперсия X.

Подставив (2.151) в (2.148), найдем

$$h(X) = \int_{-\infty}^{\infty} w(x) \log \frac{1}{w(x)}\,dx = \int_{-\infty}^{\infty} w(x)\left[\log \sqrt{2\pi\sigma^2} + \frac{\log e}{2\sigma^2}(x-a)^2\right]dx =$$

$$= \log \sqrt{2\pi\sigma^2} \int_{-\infty}^{\infty} w(x)\,dx + \frac{\log e}{2\sigma^2} \int_{-\infty}^{\infty} (x-a)^2 w(x)\,dx.$$

Первый интеграл по общему свойству плотности вероятности равен 1, а второй интеграл по определению дисперсии равен σ^2. Окончательно

$$h(X) = \log \sqrt{2\pi\sigma^2} + \frac{1}{2} \log e = \log \sqrt{2\pi\sigma^2} + \log \sqrt{e} = \log \sqrt{2\pi e \sigma^2}. \tag{2.152}$$

Таким образом, дифференциальная энтропия гауссовской случайной величины не зависит от ее математического ожидания и монотонно возрастает с увеличением дисперсии.

В заключение докажем одно важное свойство нормального распределения: из всех непрерывных случайных величин X с одинаковой дисперсией σ^2 наибольшую дифференциальную энтропию имеет величина с нормальным распределением.

Для доказательства отметим, прежде всего, что для любой плотности вероятности $w(x)$ с дисперсией σ^2 справедливо равенство

$$\int_{-\infty}^{\infty} w(x) \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2} \right) \right] dx = -\log \sqrt{2\pi e \sigma^2}.$$

Это равенство выводится совершенно таким же образом, как и (2.152). Рассмотрим теперь разность

$$h(X) - \log \sqrt{2\pi e \sigma^2} = \int_{-\infty}^{\infty} w(x) \log \frac{1}{w(x)} dx +$$

$$+ \int_{-\infty}^{\infty} w(x) \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2} \right) \right] dx =$$

$$= \log e \int_{-\infty}^{\infty} w(x) \ln \left[\frac{1}{w(x)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2} \right) \right] dx \leqslant$$

$$\leqslant \log e \int_{-\infty}^{\infty} w(x) \left[\frac{\exp\left(-\frac{x^2}{2\sigma^2} \right)}{w(x)\sqrt{2\pi\sigma^2}} - 1 \right] dx = \log e\,(1-1) = 0, \tag{2.153}$$

причем равенство имеет место только при $w(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, что и требовалось доказать.

При выводе (2.153) использовалось неравенство (2.124) и то, что интеграл от любой плотности вероятностей в бесконечных пределах равен 1.

Такую же теорему можно доказать и для многомерных гауссовских величин.

ГЛАВА ТРЕТЬЯ

КАНАЛЫ СВЯЗИ И ИХ ПРОПУСКНАЯ СПОСОБНОСТЬ

3.1. ОБЩИЕ СВЕДЕНИЯ О КАНАЛАХ СВЯЗИ

В гл. 1 канал связи определен как совокупность средств, предназначенных для передачи сообщений (под «средством» понимаются и технические устройства, и линия связи — физическая среда, в которой распространяется сигнал между пунктами связи). Канал можно представить как последовательное соединение устройств, выполняющих различные функции в общей системе связи. Такая цепь, применительно к дискретной системе радиосвязи, показана на рис. 1.5. При необходимости анализа отдельных блоков схему рис. 1.5 можно детализировать. Например, можно учесть ряд устройств, размещенных в промежуточных пунктах линии связи между передатчиком и приемником (усилительную аппаратуру вдоль линий дальней проводной связи или ретрансляторы радиорелейных линий связи).

В данном курсе не рассматриваются вопросы электрического сопряжения отдельных блоков, образующих канал связи. В первую очередь изучается выбор функциональных блоков канала, при котором была бы обеспечена наибольшая эффективность передачи информации. В зависимости от решаемых задач под каналом связи можно понимать различную совокупность блоков (см. рис. 1.3), которая в ходе решения считается заданной (см. § 1.2).

Классификация каналов связи возможна с использованием различных признаков. В зависимости от назначения систем каналы связи делят на телефонные, телевизионные, телеграфные, фототелеграфные, звукового вещания, телеметрические, смешанные и т. п. В зависимости от того, распространяется ли сигнал между пунктами связи в свободном пространстве или по направляющим линиям, выделяют каналы радио- и проводной связи (воздушные и кабельные линии связи, волноводные СВЧ тракты и т. п.).

Более существенна классификация каналов электрической связи по диапазону используемых ими частот. Так, на современных симметричных кабельных линиях связи применяют сигналы, занимающие полосы частот в диапазоне, ограниченном сверху частотой в несколько сотен килогерц. Дополнительные мероприятия по увеличению симметрии кабельных пар позволяют увеличить верхний предел используемого диапазона частот до тысяч килогерц. Коаксиальные кабели, являющиеся основой сетей магистральной дальней связи, пропускают в настоящее время диапазон частот до десятков мегагерц. На воздушных проводных линиях используются частоты не выше 150 кГц, ибо на более высо-

ких частотах в этих линиях сильно сказывается мешающее действие аддитивных помех и резко возрастает затухание в линии.

Радиосвязь осуществляется с помощью электромагнитных волн, распространяющихся в частично ограниченном (например, землей и ионосферой) пространстве. В настоящее время в радиосвязи применяются частоты примерно от $3 \cdot 10^3$ до $3 \cdot 10^{11}$ Гц. Этот огромный диапазон принято в соответствии с десятичной классификацией подразделять следующим образом:

Диапазон волн	Длины волн	Частоты
Декакилометровые (сверхдлинные, СДВ)	10—100 км	3—30 кГц
Километровые (длинные, ДВ)	1—10 км	30—300 кГц
Гектаметровые (средние, СВ)	100—1000 м	0,3—3 МГц
Декаметровые (короткие, КВ)	10—100 м	3—30 МГц
Метровые (ультракороткие, УКВ)	1—10 м	30—300 МГц
Дециметроные (ультракороткие, УКВ)	0,1—1 м	300—3000 МГц
Сантиметровые (ультракороткие, УКВ)	1—10 см	3—30 ГГц
Миллиметровые (ультракороткие, УКВ)	1—10 мм	30—300 ГГц

Примечание. В скобках указаны нестандартные, но широко используемые названия диапазонов волн.

Диапазон миллиметровых волн уже вплотную подходит к диапазону инфракрасных световых волн.

В настоящее время благодаря созданию и внедрению в практику квантовых генераторов (лазеров и мазеров) осваивается и диапазон световых волн (оптический диапазон).

Для современного этапа развития техники связи характерна тенденция к переходу на все более высокие частоты. Это вызвано рядом веских причин, в частности, необходимостью повышать скорость передачи информации, возможностью получить остронаправленное излучение при небольших размерах излучателей, меньшей интенсивностью атмосферных и многих видов промышленных помех в более высокочастотных диапазонах, возможностью применения помехоустойчивых широкополосных систем модуляции и т. д.

Наибольший интерес для теории передачи сигналов представляет классификация каналов связи по характеру сигналов на входе и выходе канала. Различают каналы:

а) дискретные (по состояниям), на входе и выходе которых сигналы дискретны. Таковы каналы, заданные между точками $a - \hat{a}$ или $b - \hat{b}$ на схеме рис. 1.2;

б) непрерывные (по состояниям), на входе и выходе которых сигналы непрерывны. Примером может служить канал, заданный между выходом модулятора и входом демодулятора в любой системе связи;

в) дискретные со стороны входа и непрерывные со стороны выхода или наоборот. Такие каналы называются дискретно-непрерывными, или полунепрерывными [например, каналы, заданные между точками $a - z$, $b - z$ (см. рис. 1.5)].

Всякий дискретный, или полунепрерывный, канал содержит внутри себя непрерывный канал. Следует помнить, что дискретность и непрерывность канала не связана с характером передаваемых сообщений. Можно передать дискретные сообщения по непрерывному каналу (см. гл. 4) и непрерывные сообщения по дискретному (см. гл. 7).

3.2. ПРОХОЖДЕНИЕ СИГНАЛОВ ЧЕРЕЗ КАНАЛЫ С ДЕТЕРМИНИРОВАННЫМИ ХАРАКТЕРИСТИКАМИ

Передача сигналов по реальным каналам связи всегда сопровождается изменениями (преобразованиями) этих сигналов, в результате чего принятые сигналы отличаются от переданных. Отличия эти обусловлены, прежде всего, линейными и нелинейными преобразованиями входных сигналов, а также наличием аддитивных шумов в канале, существующих чаще всего независимо от передаваемых сигналов. С точки зрения передачи информации по каналу, важно подразделение преобразований сигнала на обратимые и необратимые. Как уже отмечалось (см. § 2.9), обратимые преобразования не влекут за собой потери информации. Это значит, что взаимная информация между сигналами на входе и выходе канала равна собственной информации входного сигнала. При необратимых преобразованиях потери информации неизбежны. Для обратимых преобразований сигнала часто используют термин «искажение», а необратимые преобразования называют помехами (аддитивными и неаддитивными).

Примером простейшего детерминированного обратимого преобразования входного сигнала $X(t)$, которое не меняет его форму, служит

$$Y(t)=kX(t-\tau). \tag{3.1}$$

В данном случае выходной сигнал канала $Y(t)$ отличается от входного лишь известным масштабом k, который легко компенсируется соответствующим усилением или ослаблением сигнала и постоянной задержкой во времени τ. Задержка сигнала во времени приводит к задержке приема информации, но не к потере ее. Она чаще всего невелика. По существу, лишь при связи в масштабах космоса или при очень большом числе реактивных элементов в линии связи задержка может оказаться ощутимой[1].

Если входной сигнал $X(t)$ в (3.1) узкополосный, его удобно представить в квазигармонической форме (2.68): $X(t)=A(t)\cos[\omega_0 t+\Phi(t)]$, где $A(t)$ и $\Phi(t)$ — медленно меняющиеся функции. Поэтому при достаточно малой задержке τ можно в

[1] Здесь идет речь о задержке в самом канале, а не о задержках в демодуляторе и декодере, которые могут быть значительными и иногда лимитируют возможность повышения помехоустойчивости.

первом приближении считать $A(t-\tau) \approx A(t)$ и $\Phi(t-\tau) \approx \Phi(t)$, а выходной сигнал в (3.1) записать следующим образом:

$$Y(t) = kA(t-\tau)\cos[\omega_0(t-\tau)+\Phi(t-\tau)] \approx kA(t)\cos[\omega_0 t+\Phi(t)-\theta_к], \tag{3.2}$$

где $\theta_к = \omega_0\tau$ — фазовый сдвиг в канале.

Таким образом, при узкополосном сигнале малая задержка сводится к некоторому сдвигу фазы.

В реальных каналах связи, даже когда можно пренебречь аддитивным шумом, преобразования сигналов имеют сложный характер и обычно приводят к отличию формы выходного сигнала от входного.

Исследование преобразований случайных процессов при их прохождении через динамические системы (как с регулярными, так и со случайно меняющимися параметрами) связано с решением задач двух типов:

1. По данной корреляционной функции (или энергетическому спектру) входного воздействия $X(t)$ найти корреляционную функцию (энергетический спектр) отклика $Y(t)$ на выходе динамической системы, заданной ее характеристиками.

2. Зная многомерные распределения входного воздействия $X(t)$, найти многомерные распределения отклика $Y(t)$ на выходе заданной динамической системы.

Вторая из указанных задач является более общей. Из ее решения, очевидно, может быть получено решение и первой задачи. Однако в дальнейшем в основном ограничимся кратким рассмотрением первой задачи и лишь укажем возможные пути решения второй, более сложной задачи.

Прохождение случайных сигналов через детерминированные линейные цепи. Как известно, линейная цепь с постоянными параметрами характеризуется своей импульсной реакцией $g(t)$ или ее преобразованием Фурье — передаточной функцией $k(i\omega)$. Если, например, на вход цепи поступает центрированный процесс $X(t)$, то процесс $Y(t)$ на выходе определяется интегралом Дюамеля[1]

$$Y(t) = \int_{-\infty}^{\infty} g(\tau)\,X(t-\tau)\,d\tau. \tag{3.3}$$

В физически реализуемой цепи $g(t) = 0$ при $t<0$. Поэтому нижний предел в (3.3) можно заменить нулем.

Найдем функцию корреляции центрированного выходного процесса $Y(t)$:

$$B_Y(t_1, t_2) = M\{Y(t_1)\,Y(t_2)\} = M\left\{\int_{-\infty}^{\infty} g(\tau)\,X(t_1-\tau)\,d\tau \int_{-\infty}^{\infty} g(\tau)\,X(t_2-\right.$$

[1] Здесь и в дальнейшем интегрирование случайных процессов понимается в среднеквадратичном смысле [см. ф-лу (2.8)].

$$-\tau)\,d\tau\Big\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} M\{X(t_1-\tau_1)X(t_2-\tau_2)\}\,g(\tau_1)\,g(\tau_2)\,d\tau_1\,d\tau_2 =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} B_X^{\text{т}}(\theta_1,\ \theta_2)\,g(t_1-\theta_1)\,g(t_2-\theta_2)\,d\theta_1\,d\theta_2,$$

где $\theta_1 = t_1 - \tau_1$; $\theta_2 = t_2 - \tau_2$; $B_X(\theta_1, \theta_2)$ — функция корреляции входного сигнала.

Пусть входной процесс стационарен. Тогда $B_X(\theta_1, \theta_2) = B_X(\vartheta)$, где $\vartheta = \theta_2 - \theta_1$. Введем также обозначения $t_2 - t_1 = \tau$, $t_1 - \theta_1 = \tau_1$. Тогда $t_2 - \theta_2 = \tau + \tau_1 - \vartheta$ и

$$B_X(t_1,\ t_1+\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} B_X(\vartheta)\,g(\tau_1)\,g(\tau+\tau_1-\vartheta)\,d\tau_1 d\tau =$$

$$= \int_{-\infty}^{\infty} B_X(\vartheta)\,B_g(\tau-\vartheta)\,d\vartheta, \qquad (3.4)$$

где использована «временна́я функция корреляции» (ВФК) от неслучайной импульсной реакции $B_g(\beta) = \int_{-\infty}^{\infty} g(\tau_1)g(\tau_1+\beta)d\tau_1$. В данном случае $\beta = \tau - \vartheta$.

Из (3.4) видно, что при стационарном входном процессе и выходной процесс оказывается стационарным, так как $B_Y(t_1, t_1+\tau)$ не зависит от t_1. Поэтому можно записать

$$B_Y(\tau) = \int_{-\infty}^{\infty} B_X(\vartheta)\,B_g(\tau-\vartheta)\,d\vartheta. \qquad (3.5)$$

Полученное равенство является аналогом интеграла Дюамеля для корреляционных функций. Таким образом, ФК выходного процесса является интегральной сверкой ФК входного процесса и ВФК импульсной реакции цепи. Заметим, что ВФК импульсной реакции связана преобразованием Фурье с квадратом модуля передаточной функции $|k(i\omega)|^2$ или амплитудно-частотной характеристики (АЧХ) цепи. Действительно,

$$|k(i\omega)|^2 = k(i\omega)\,k^*(i\omega) = \int_{-\infty}^{\infty} g(t)\,e^{-i\omega t}\,dt \int_{-\infty}^{\infty} g(t)\,e^{i\omega t}\,dt =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t_2)\,e^{-i\omega t_2} g(t_1)\,e^{i\omega t_1}\,dt_1 dt_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t_1)\,g(t_1+\beta)\,e^{-i\omega\beta}\,dt_1 d\beta =$$

$$= \int_{-\infty}^{\infty} B_g(\beta)\,e^{-i\omega\beta}\,d\beta. \qquad (3.6)$$

Из теории преобразования Фурье известно, что преобразование Фурье от свертки двух функций равно произведению преобразований Фурье от этих функций. Применив это к (3.5), получим простое соотношение между спектральными плотностями стационар-

ных процессов на входе и на выходе линейной цепи с постоянной передаточной функцией $k(i\omega)$:

$$G_Y(f) = G_X(f) \mid k(i\,2\pi f)\mid^2. \tag{3.7}$$

Из (3.5) и (3.7) следует, что ФК и спектр процесса на выходе цепи полностью определяются ФК или спектром процесса на входе и АЧХ цепи, т. е. не зависят ни от распределения вероятностей входного процесса, ни от фазо-частотной характеристики цепи.

При подаче сигнала $X(t)$ на детерминированную линейную цепь с переменными параметрами выходной сигнал $Y(t)$, как известно, можно выразить интегралом свертки:

$$Y(t) = \int_{-\infty}^{\infty} g(t,\ \tau)\, X(t-\tau)\, d\tau, \tag{3.8}$$

где $g(t, \tau)$ — функция двух переменных, определяющая реакцию системы в момент t на δ-импульс, поданный на вход в момент $t-\tau$.

Преобразование Фурье от $g(t, \tau)$ по переменной τ

$$k(t,\ i\omega) = \int_{-\infty}^{\infty} g(t,\ \tau)\, e^{-i\omega\tau} d\tau \tag{3.9}$$

представляет передаточную функцию линейной цепи с переменными параметрами, которая, естественно, является функцией не только частоты, но и времени.

Поскольку в физически реализуемой цепи отклик не может возникнуть раньше воздействия, $g(t, \tau) = 0$ при $\tau < 0$. Поэтому в (3.8) и (3.9) нижний предел можно заменить нулем.

Пусть на вход линейной цепи с передаточной функцией (3.9) поступает стационарный сигиал $X(t)$ со спектральной плотиостью мощности $G_X(f)$.

Нетрудно показать, что ФК выходного сигнала равна обратному преобразованию Фурье от $G_X(f)B_c(f, t, \tau)$:

$$B_Y(t,\ t+\tau) = \int_{-\infty}^{\infty} G_X(f)\, B_c(f,\ t,\ \tau) \exp(i\,2\pi f\tau)\, df, \tag{3.10}$$

где $B_c(f, t, \tau) = k(-i\omega, t)k(i\omega, t+\tau)$ — характеристика системы с переменными параметрами [12].

Заметим, что сигнал $Y(t)$ на выходе цепи с переменными параметрами в общем случае является нестационарным процессом, даже если входной сигнал $X(t)$ стационарен.

Уравнение (3.3) является частным случаем уравнения (3.8), если в линейной системе параметры могут считаться неизменными во времени. В этом случае реакция $g(t, \tau)$ зависит лишь от разности моментов наблюдения отклика t и подачи воздействия $t-\tau$: $g(t, \tau) = g(\tau)$.

Другой простой частный случай линейной цепи имеет место, когда импульсная реакция $g(t, \tau) = g_1(t)\delta(\tau)$. Эта цепь является

безынерционной и представляет собой перемножитель. Действительно,

$$Y(t)=\int_{-\infty}^{\infty} g_1(t)\,\delta(\tau)\,X(t-\tau)\,d\tau=g_1(t)\,X(t). \tag{3.11}$$

В несколько более общем случае, когда $g(t,\tau)=g_1(t)g_2(\tau)$, линейная цепь с переменными параметрами сводится к каскадному соединению перемножителя и цепи с постоянными параметрами, так как

$$Y(t)=\int_{-\infty}^{\infty} g_1(t)\,g_2(\tau)\,X(t-\tau)\,d\tau=g_1(t)\int_{-\infty}^{\infty} g_2(\tau)\,X(t-\tau)\,d\tau. \tag{3.12}$$

Приведем еще один пример: цепь с переменной задержкой, в которой $g(t,\tau)=a\delta[\tau-\Delta(t)]$, где a — постоянный масштабный множитель; $\Delta(t)$ — некоторая известная функция времени. Выходной сигнал

$$Y(t)=a\int_{-\infty}^{\infty}\delta(\tau-\Delta(t))\,X(t-\tau)\,d\tau=aX(t-\Delta(t)). \tag{3.13}$$

Передаточная функция этой цепи

$$k(t,\ i2\pi f)=\int_{-\infty}^{\infty} a\,\delta(\tau-\Delta(t))\,e^{-i2\pi f\tau}\,d\tau=a\exp[-i2\pi f\Delta(t)]. \tag{3.14}$$

Корреляционная функция выхода согласно (3.10)

$$B_Y(t,\ \tau)=a^2\int_{-\infty}^{\infty} G_X(f)\exp\{i2\pi f[\tau+\Delta(t)-\Delta(t+\tau)]\}\,df. \tag{3.15}$$

Очевидно, что интеграл выражает корреляционную функцию процесса $X(t)$ при аргументе $\tau+\Delta(t)-\Delta(t+\tau)$, следовательно,

$$B_Y(t,\ \tau)=a^2B_X[\tau+\Delta(t)-\Delta(t+\tau)]. \tag{3.16}$$

Если система обеспечивает неизменную во времени задержку $\Delta(t)=\text{const}$, то $\Delta(t)-\Delta(t+\tau)=0$ и корреляционная функция выходного сигнала с точностью до постоянного множителя a^2 равна корреляционной функции входного.

Рассмотрим случай, когда задержка линейно зависит от времени, т. е. $\Delta(t)=\kappa t$, где κ — коэффициент пропорциональности.

Такая задержка возникает, например, в системе радиосвязи или акустической связи, если точки приема и передачи находятся в относительном движении (например, при связи через спутник из-за перемещения спутника относительно антенны приемника). В этом случае $\kappa=v_r/c$, где c — скорость распространения сигнала в среде, v_r — радиальная составляющая скорости взаимного движения передатчика относительно приемника, причем v_r имеет положительный знак при их взаимном удалении и отрицательный при сближении. Обычно $|\kappa|\ll 1$. Вместо (3.16) можно теперь записать

$$B_Y(\tau)=a^2B_X[(1-k)\tau], \tag{3.17}$$

т. е. при стационарном воздействии процесс на выходе четырехполюсника остается стационарным.

Корреляционной функции (3.17) соответствует энергетический спектр

$$G_Y(f)=a^2\int_{-\infty}^{\infty} B_X[(1-k)\tau]\exp(-i2\pi f\tau)\,d\tau=\frac{a^2}{1-k}\,G_X\left(\frac{f}{1-k}\right)\approx$$

$$\approx\frac{a^2}{1+k}\,G_X(f+kf). \tag{3.18}$$

Таким образом, линейное изменение задержки приводит к смещению средней частоты спектра f_0 на величину κf_0 (*допплеровскому смещению частоты*) и, кроме того, — сужению (при $\kappa>0$) или расширению (при $\kappa<0$) энергетического спектра процесса, а также изменению масштабного множителя. В обычных условиях, когда $|\kappa|\ll 1$, только первый из этих эффектов может иметь существенное значение.

Задача нахождения распределения вероятностей отклика линейной системы при произвольном случайном воздействии оказывается в общем случае весьма сложной, даже если ограничится нахождением одномерного распределения. Отметим, однако, что если на вход линейной детерминированной системы подан гауссовский процесс, то и процесс на выходе оказывается гауссовским, что следует из известных свойств нормального распределения, которое остается нормальным при любых линейных преобразованиях. Если процесс на входе не гауссовский, то при прохождении линейной системы его распределение вероятностей меняется иногда весьма существенно.

Отметим общее свойство, присущее узкополосным линейным системам. Если полоса частот F_c, занимаемая входным сигналом $X(t)$, много шире полосы пропускания данной линейной системы F, то распределение выходного процесса имеет тенденцию приближаться к нормальному. Это можно грубо пояснить, исходя из формулы (3.8). Узость полосы пропускания означает, что длительность импульсной реакции $g(t, \tau)$ как функции τ велика по сравнению с интервалом корреляции входного процесса $X(t)$. Поэтому сечение выходного процесса $Y(t)$ в любой момент t определяется интегралом (3.8), в подынтегральную функцию которого с достаточно большим весом входит много некоррелированных между собой сечений процесса $X(t)$. Распределение вероятностей такого интеграла согласно центральной предельной теореме должно быть близким к нормальному, тем ближе, чем больше отношение ширины спектра входного сигнала к полосе пропускания цепи. В предельном случае, если на вход цепи воздействует белый шум, у которого ширина спектра бесконечна, а цепь имеет ограниченную полосу пропускания, то выходной процесс будет гауссовским.

Прохождение случайных сигналов через нелинейные цепи. Ограничимся рассмотрением только безынерционных нелинейных систем с регулярными параметрами, у которых вход и выход связаны некоторой нелинейной зависимостью, называемой характеристикой системы:

$$y(t)=\varphi[x(t)]. \tag{3.19}$$

Соотношением вида (3.19) достаточно точно может быть охарактеризована работа ряда звеньев реальных каналов связи, например входящих в состав демодуляторов, ограничителей, модуляторов и т. п. Преобразование $x(t)\to y(t)$, как правило, однозначно, что не всегда можно сказать об обратном преобразовании $y(t)\to x(t)$ (например, квадратичная цепь с характеристикой $y=kx^2$). В силу неприменимости суперпозиции к нелинейным системам рассмотрение сложного воздействия (например, суммы детермини-

рованного и случайного слагаемых) нельзя свести к рассмотрению прохождения каждого из компонентов в отдельности.

При нелинейных преобразованиях возникает трансформация (изменение) спектра входного воздействия. Так, если на вход нелинейной системы воздействует смесь регулярного сигнала и аддитивного шума $X(t)=u(t)+N(t)$ в узкой полосе частот F_c, группирующейся около средней частоты f_0, то в общем случае на выходе будут присутствовать составляющие комбинационных частот трех видов, группирующиеся около частот $nf_0(n=0, 1, 2)$: продукты биений составляющих входного сигнала между собой (с×с), продукты биений составляющих входного шума (ш×ш); продукты биений сигнала и шума (с×ш). Разделить их на выходе системы обычно невозможно.

Если известны характеристика $y=\varphi(x)$ нелинейной системы и двумерная функция распределения входного воздействия $w(x_1, x_2; t_1, t_2)$, то статистические характеристики выходного процесса, в принципе, всегда можно определить. Так, математическое ожидание отклика

$$\overline{Y(t)}=\overline{\varphi(X(t))}=\int_{-\infty}^{\infty}\varphi(x)\,w(x)\,dx, \tag{3.20}$$

а его корреляционная функция

$$B_Y(t,\ \tau)=\int_{-\infty}^{\infty}[\varphi(x_1(t))-\overline{Y(t)}]\,[\varphi(x_2(t+\tau))-$$

$$-\overline{Y(t+\tau)}]\,w(x_1,\ x_2;\ t,\ t+\tau)\,dx_1dx_2. \tag{3.21}$$

Обратным преобразованием Фурье можно по (3.21) найти и энергетический спектр.

Используя правила нахождения законов распределения для функций от случайных величин (случайных процессов), можно, в принципе, находить и распределение выходного процесса любого порядка, если известно распределение входного процесса. Однако определение вероятностных характеристик отклика нелинейных систем (цепей) даже на стационарные входные воздействия оказывается весьма громоздким и сложным, несмотря на то, что для решения этой задачи разработан ряд специальных приемов. Во многих случаях, особенно для узкополосных сигналов, эти расчеты существенно упрощаются при использовании квазигармонического представления процесса.

3.3. СЛУЧАЙНЫЕ ПРЕОБРАЗОВАНИЯ СИГНАЛОВ В КАНАЛАХ СВЯЗИ

Помимо рассмотренных в предыдущем параграфе детерминированных преобразований сигнала, в отдельных звеньях канала (в частности, в линии связи) имеют место и случайные преобразования сигнала. В простейшем случае это преобразование сводится к

суммированию сигнала с независимым от него случайным процессом, называемым аддитивной помехой или аддитивным шумом. В более сложных каналах к этому добавляются случайные изменения параметров канала, в результате которых даже в отсутствие аддитивных помех принимаемый сигнал не определяется однозначно передаваемым.

Рассмотрим в общих чертах характерные преобразования сигнала в случайных линейных и нелинейных цепях, а также в цепях с аддитивной помехой.

Случайные линейные цепи. В самом общем виде линейная цепь описывается случайной импульсной реакцией $G(t, \tau)$, имеющей тот же смысл, что и $g(t, \tau)$ в (3.8), но представляющей случайную функцию двух аргументов, t (момента наблюдения реакции) и τ (времени прошедшего с момента подачи δ-импульса на вход цепи). Такова, например, импульсная реакция любой линейной цепи, параметры которой подвергаются воздействию случайных внешних влияний, например температуры, давления, влажности и т. д.

Случайную линейную цепь можно характеризовать также случайной передаточной функцией переменных t и ω:

$$K(t, i\omega) = \int_{-\infty}^{\infty} G(t, \tau) e^{-i\omega\tau} d\tau.$$

Остановимся на нескольких частных случаях, с которыми чаще всего приходится встречаться. Простейшим является случайный канал, описываемый уравнением (3.1), если его параметры k и τ флуктуируют. Обычно такие флуктуации, например, в различных проводных линиях связи вызываются изменениями внешних условий и происходят чрезвычайно медленно [1] и в очень небольших относительных пределах. В радиоканалах при прямом распространении, гидроакустических каналах и т. п. флуктуации выражены несколько более заметно.

Из уравнения (3.2) видно, что при передаче по такому каналу узкополосных сигналов с достаточно высокой средней частотой флуктуации времени задержки приводят к флуктуации начальной фазы $\theta_к$ сигнала на выходе канала. Важно отметить, что даже при очень малых относительных флуктуациях времени задержки τ начальная фаза может изменяться в очень больших пределах. Для этого необходимо выполнить условие $\Delta\tau \gg 1/f_0$, где $\Delta\tau$ — среднее квадратичное отклонение задержки; f_0 — средняя частота сигнала. Это условие в радиоканалах обычно выполняется.

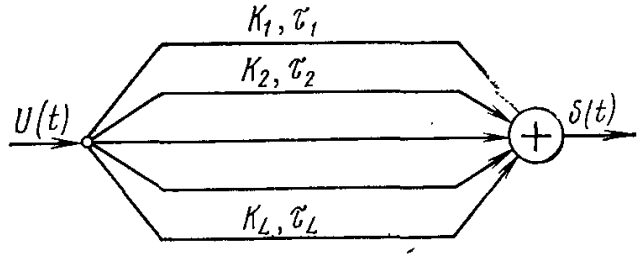


Рис. 3.1. Многопутевое распространение сигнала

Более сложный случай имеет место, когда сигнал проходит по нескольким параллельным путям от входа канала к его выходу (рис. 3.1), так что на выходе каждого пути сигнал имеет вид (3.1), но значения k и τ для разных путей различны и к тому же в небольших пределах флуктуируют.

Такого рода многопутевое распространение сигнала характерно для большинства радио-, гидроакустических и некоторых других каналов. Энергия волны распространяется обычно в неоднородной среде и испытывает отражения от различных неоднородностей. Эти неоднородности могут быть распределены внутри относительно небольшого отражающего (рассеивающего) объема. В этом случае разности хода (разности значений τ) для отдельных путей невелики.

[1] Это значит, что за время длительности «котельниковского интервала» $1/(2F)$, где F — ширина спектра сигнала, параметры канала не успевают заметно измениться.

Если по такому каиалу направить очень короткий импульс, то и на его выходе импульс будет довольно коротким. Такой канал принято называть однолучевым. Наличие разных путей («подлучей», как их иногда называют) не вызывает в этом случае существенного рассеяния (растяжения) сигнала во времеии, но приводит к возникновению явления *замираний*, которое заключаетси в более или менее быстрых случайных изменениях передаточной фуикции канала.

Для пояснения замираний рассмотрим передачу по каналу (см. рис. 3.1) гармоннческого сигнала с единичной амплитудой $u(t)=\mathrm{Re}\, e^{i\omega t}$. На выходе получим сигнал

$$S(t)=\mathrm{Re}\sum_{l=1}^{L}K_l e^{-i\omega(t-\tau_l)}=\mathrm{Re}\,\dot{K}e^{i\omega t}, \tag{3.22}$$

где L — число путей (подлучей, попадающих в точку приема); K_l — коэффициент передачи по l-му подлучу; τ_l — время распространения l-го подлуча; $\dot{K}=\sum_{i=1}^{L}K_l e^{-i\omega\tau_l}$ комплексная амплитуда выходного сигиала, которая в даином случае по определению равна передаточной функции канала.

Передаточная функция в общем случае зависит от частоты ω. Если же учесть, что вследствие хаотических перемещений отражателей значения K_l н τ_l флуктуируют, то $\dot{K}$ зависит также от времени, представляя собой случайную функцию (мультипликативную помеху) $\dot{K}(t, i\omega)$. Во многих случаях передаточная функция $\dot{K}(t, i\omega)$ флуктуирует зиачительно быстрее, чем величиины K_l и τ_l.

Важной характернстикой канала с замираниями является распределение вероятностей комплексной передаточной функции $\dot{K}(t, i\omega)$ и, в первую очередь, ее модуля $|\dot{K}|$. Для ее определения представим $\dot{K}(t, i\omega)$ в следующем виде:

$$\dot{K}(t, i\omega)=Ke^{-i\theta_k}=X(t, \omega)+iY(t, \omega), \tag{3.23}$$

где $K=|\dot{K}|$ и θ_k — модуль и аргумент передаточной функции, которые также являются случайными функциями t и ω, а $X=K\cos\theta_k$ и $Y=K\sin\theta_k$ — квадратурные составляющие.

С другой стороны, согласно (3.22),

$$\dot{K}(t, i\omega)=\sum_{l=1}^{L}K_l e^{-i\omega\tau_l}=\sum_{l=1}^{L}K_l\cos\omega\tau_l-i\sum_{l=1}^{L}K_l\sin\omega\tau_l,$$

откуда

$$X(t, \omega)=\sum_{l=1}^{L}K_l\cos\omega\tau_l;\quad Y(t, \omega)=\sum_{l=1}^{L}K_l\sin\omega\tau_l. \tag{3.24}$$

Поскольку X и Y образуются в результате сложения большого числа слабо коррелированных величин с ограниченными дисперсиямн, к ним обычно можио применить центральную предельную теорему н счнтать их нормально распределенными.

Для случая, когда все K_l одного порядка и фазовые сдвиги достаточио велики, легко показать, что X и Y имеют одинаковые дисперсии σ^2, а их математические ожндания $m_x=m_y=0$. В этом случае одномерное распределение вероятностей $|K|$ является рэлеевским:

$$\left.\begin{aligned} w(k)&=\frac{k}{\sigma^2}\exp\left(-\frac{k^2}{2\sigma^2}\right), \quad k\geqslant 0;\\ w(k)&=0, \qquad\qquad\qquad k<0. \end{aligned}\right\} \tag{3.25}$$

Это доказывается в точности так, как было сделано в § 2.4 при выводе формулы (2.76). Фаза результирующего сигнала θ_k при этом распределена равномерно на интервале $(0; 2\pi)$. Дисперсия квадратурных составляющих σ^2 равна средней мощности приходящего сигнала. Такие замирания, как и каналы, в которых они проявляются, называются *рэлеевскими.*

Во многих каналах замирания отличаются от рэлеевских. Иногда в одном из подлучей коэффициент передачи k_l значительно больше, чем в других, и можно считать, что помимо диффузно отраженных подлучей в место приема приходит и регулярный (незамирающий) луч. В этом случае коэффициент передачи канала $K=\sqrt{X^2+Y^2}$ подчиняется обобщенному распределению Рэлея [см. (2.86)]:

$$w(k)=\frac{k}{\sigma^2}\exp\left(-\frac{k^2}{2\sigma^2}-q^2\right)I_0\left(\frac{k\sqrt{2}q}{\sigma}\right),\quad k\geqslant 0. \tag{3.26}$$

Здесь $q^2=(m^2_x+m^2_y)/2\sigma^2$ — отношение мощностей регулярной и флуктуирующей составляющих.

В общем случае, когда $\sigma^2_x\neq\sigma^2_y$, получается так называемое четырехпараметрическое распределение модуля и фазы замирающего сигнала. Соответствующие плотности вероятностей даны в [8].

Если по однолучевому каналу с замираниями передается относительно узкополосный сигнал, а среднее квадратичное отклонение запаздывания $\Delta\tau$ в отдельных подлучах удовлетворяет условию

$$\Delta\tau\ll 1/F, \tag{3.27}$$

где F — ширина спектра сигнала, то изменения начальных фаз на разных частотах ω в спектре сигнала, равные $\omega\Delta\tau$, почти одинаковы. При этом все составляющие спектра сигнала замирают «дружно», т. е. их амплитуда и фазы изменяются одинаково. Такие замирания называют общими, или гладкими [1]. Если же условие (3.27) не выполнено, то в разных областях спектра сигнала процессы замирания не совпадают *(селективные замирания).* При этом наблюдаются существенные изменения формы сигнала.

Быстрота изменений во времени комплексного случайного процесса $K(t, i\omega)=X(t, \omega)+iY(t, \omega)$ (при фиксированной частоте), или, как говорят, скорость замираний сигнала характеризуется временем автокорреляции τ_k квадратурных компонент $X(t, \omega)$ и $Y(t, \omega)$ или шириной энергетического спектра замираний $\Delta f_{зам}=1/\tau_k$.

По экспериментальным данным для ионосферной коротковолновой радиосвязи $\tau_k=0{,}1—1$ с и более. Для других каналов значения τ_k могут меняться в широких пределах. При исследовании условий передачи сигналов в каналах с замираниями существенна не величина $\tau_к$ сама по себе, а ее отношение к длительности элемента сигнала T* или отношение скорости замираний $1/\tau_к$ к скорости передачи сигналов $1/T$. Интерференционные замирания называют медленными, если $\tau_к\gg T$, что справедливо для большинства радиоканалов. В этом случае можно считать, что X и Y (K и θ_k) не меняются на интервале T — длительности элемента сигнала. Если передача идет с очень большими скоростями, когда $T/\tau_к\to 0$, можно считать, что практически X и Y остаются неизменными на протяжении многих элементов сигнала. Если $\tau_к$ одного порядка с T и даже меньше, замирания называют быстрыми.

Явление, напоминающее глубокие замирания, но имеющие совершенно другую природу, нередко наблюдается в кабельных каналах в виде кратковременных (с длительностью порядка миллисекунд) прерываний связи, когда сигнал полностью или частично пропадает. Эти прерывания вызваны коммутационными процессами. Они практически не влияют на телефонную связь, но оказываются весьма существенными при передаче дискретных сообщений с

---

[1] Заметим, что условие (3.27) может выполняться при $\Delta\tau\gg 1/f_{01}$, так как в радиоканалах $f_0\gg F$.

* При передаче непрерывных сообщений под T следует понимать котельниковский интервал $1/(2F)$.

большой скоростью, поскольку каждый перерыв препятствует приему нескольких символов.

Нелинейные преобразования в канале. Среда, в которой распространяются радиосигналы, а также проводные линии связи обладают высокой степенью линейности. Почти все нелинейные искажения, наблюдаемые в некоторых каналах, вызываются входящей в состав канала аппаратурой, в частности выходными каскадами передатчика, входными каскадами приемника и промежуточными усилителями (ретрансляторами).

Промежуточные усилители можно разделить на два типа — линейные (аналоговые) и нелинейные (цифровые). И те, и другие применяются как в проводной, так и в радиорелейной связи. Аналоговые ретрансляторы проектируются так, чтобы обеспечить весьма высокую степень линейности. Это достигается применением глубокой отрицательной обратной связи в усилителях, позволяющей снизить коэффициент гармоник до сотых долей процента. Это очень существенно для широкополосных каналов, в которых передается одновременно много сообщений с частотным разделением (см. гл. 8), так как нелинейность в этом случае приводит к взаимным (переходным) помехам.

В каналах, называемых цифровыми, все или часть ретрансляторов содержат демодулятор и модулятор, которые, как правило, нелинейны по самому принципу действия. На каждом участке между ретрансляторами сигнал искажается аддитивными и мультипликативными помехами, но если длина участка невелика, эти искажения не настолько значительны, чтобы привести к ошибочной демодуляции. Поэтому демодулятор ретранслятора восстанавливает с вероятностью, очень близкой к 1, дискретный сигнал, действовавший на входе предыдущего модулятора. Это позволяет вновь сформировать непрерывный сигнал таким же, каким он был сформирован в передающем устройстве. Такие нелинейные ретрансляторы называют регенераторами (восстановителями).

Если канал с линейными ретрансляторами (рис. 3.2*а*) можно рассматривать как последовательное соединение нескольких непрерывных каналов, канал

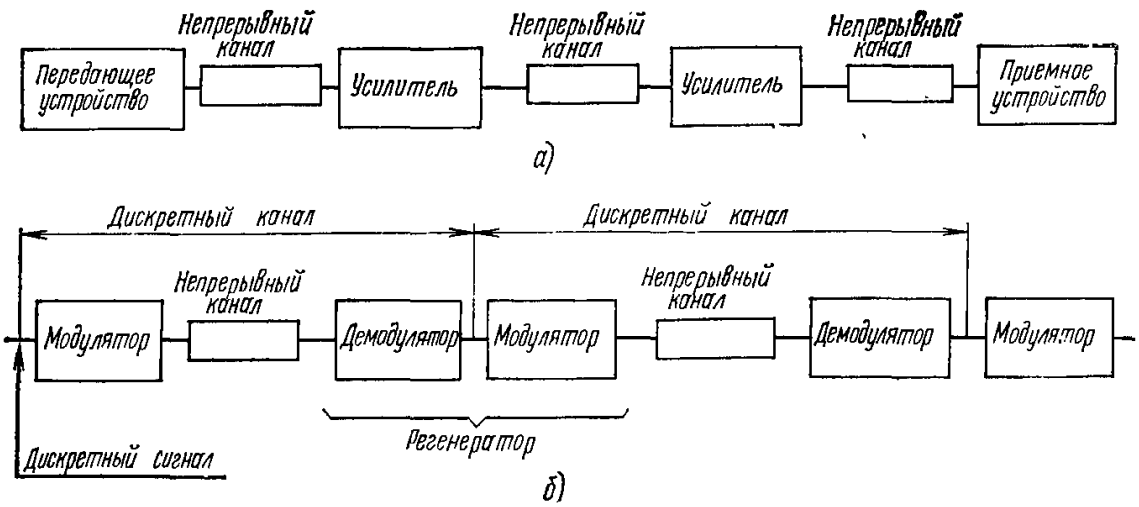


Рис. 3.2. Структурная схема канала связи с ретрансляторами:
а) линейными; *б*) нелинейными (регенераторами)

с регенераторами (рис. 3.2*б*) является последовательным соединением нескольких дискретных каналов. В состав каждого из них входят модулятор, непрерывный канал (среда, линия) и демодулятор.

Аддитивные помехи. Помехи в канале связи вызываются весьма различными причинами и могут принимать самые различные формы, индивидуальные реализации которых трудно учесть. Именно эти помехи чаще вызывают необратимые преобразования передаваемых сигналов. Как уже указывалось в гл. 1, несмотря

на большое разнообразие, аддитивные помехи по их электрической и статистической структуре разделяют на три основных класса: флуктуационные (распределенные по частоте и времени), сосредоточенные по частоте (квазигармонические) и сосредоточенные во времени (импульсные). Следует еще раз подчеркнуть, что с точки зрения передачи информации по каналу существен не вид преобразования сигнала, а то, являются ли эти преобразования обратимыми или необратимыми. Последние влекут за собой потери информации.

В отсутствие аддитивных помех детерминированные линейные преобразования сигнала чаще всего обратимы. В присутствии даже очень слабой аддитивной помехи линейные преобразования оказываются необратимыми.

Нелинейные безынерционные детерминированные преобразования сигнала необратимы в том случае, когда значения входного сигнала являются неоднозначной функцией от выходного сигнала. Так, например, преобразование $S(t)=kx^2(t)$ необратимо, так как по $S(t)$ можно восстановить $x(t)$ только с точностью до знака. Примерами необратимых нелинейных преобразований являются амплитудное детектирование (нельзя восстановить мгновенную частоту и фазу входного сигнала), идеальное амплитудное ограничение (нельзя восстановить огибающую) и т. п. Случайные помехи, как аддитивные, так и мультипликативные, приводят к необратимым преобразованиям.

Всякое обратимое преобразование сигнала (искажение), в принципе, можно скомпенсировать, включив на выходе канала цепь, создающую обратное искажение. Это называется коррекцией канала, которая довольно часто используется, особенно в проводной связи, где преобразования, как правило, линейны, детерминированны и обратимы. Частичная коррекция возможна и при случайных изменениях в канале, если они достаточно медленны и имеется возможность их проанализировать и на некотором интервале времени считать детерминированными. Такая коррекция называется адаптивной. Не следует, однако, считать, что применение коррекции канала является обязательным при обратимых преобразованиях. Поскольку обратимые преобразования не влекут за собой потери информации, можно и из искаженного сигнала, не прибегая к коррекции, извлечь информацию о переданном сообщении, что собственно и является задачей системы связи. Коррекция применяется в тех случаях, когда она облегчает последующую обработку сигнала.

3.4. МАТЕМАТИЧЕСКИЕ МОДЕЛИ КАНАЛОВ СВЯЗИ

Для того чтобы дать математическое описание канала, необходимо и достаточно указать множество сигналов, которые могут быть поданы на его вход, и для любого допустимого входного сигнала задать случайный процесс (сигнал) на выходе канала. Задание процесса понимается в том смысле, как это было определено

в § 2.1, и сводится к заданию в той или иной форме распределения вероятностей.

Точное математическое описание любого реального канала обычно оказывается весьма сложным. Вместо этого пользуются упрощенными математическими моделями, которые позволяют выявить все важнейшие закономерности реального канала, если при построении модели учтены наиболее существенные особенности канала и отброшены второстепенные детали, мало влияющие на ход связи.

Рассмотрим наиболее простые и широко используемые математические модели каналов, начав с непрерывных каналов, поскольку они во многом предопределяют и характер дискретных каналов.

Идеальный канал без помех представляет собой линейную цепь с постоянной передаточной функцией, обычно сосредоточенной в ограниченной полосе частот. Допустимы любые входные сигналы, спектр которых лежит в определенной полосе частот F и имеющие ограниченную среднюю мощность P_c (либо пиковую мощность $P_{пик}$). Эти ограничения характерны для всех непрерывных каналов, и в дальнейшем они оговариваться не будут. Заметим, что если мощность сигнала не ограничивать, но считать конечной, то множество допустимых сигналов образует векторное пространство, конечномерное (при определенных ограничениях на длительность и ширину спектра) либо бесконечномерное (при более слабых ограничениях). В идеальном канале выходной сигнал при заданном входном оказывается детерминированным. Эта модель иногда используется для описания кабельных каналов. Однако, строго говоря, она непригодна для реальных каналов, в которых неизбежно присутствуют, хотя бы и очень слабые, аддитивные помехи.

Канал с аддитивным гауссовским шумом, в котором сигнал на выходе

$$Z(t) = ku(t-\tau) + N(t), \tag{3.28}$$

где $U(t)$ — входной сигнал; k и τ — постоянные; $N(t)$ — гауссовский аддитивный шум с нулевым математическим ожиданием и заданной корреляционной функцией. Чаще всего рассматривается белый шум либо квазибелый (с равномерной спектральной плотностью в полосе спектра сигнала $U(t)$).

Обычно запаздывание τ не учитывают, что соответствует изменению начала отсчета времени на выходе канала.

Некоторое усложнение этой модели получается, если коэффициент передачи k и запаздывание τ считать известными функциями времени:

$$Z(t) = k(t)U[t-\tau(t)] + N(t). \tag{3.29}$$

Такая модель удовлетворительно описывает многие проводные каналы, радиоканалы при связи в пределах прямой видимости, а

также радиоканалы с медленными общими замираниями, при которых можно надежно предсказать значения k и τ.

Канал с неопределенной фазой сигнала отличается от предыдущего тем, что в нем запаздывание является случайной величиной. Для узкополосных сигналов, с учетом (2.69) и (3.2), выражение (3.29) при постоянном k и случайных $\tau(t)$ можно представить в виде

$$Z(t) = k[u(t)\cos\theta_k + \tilde{u}(t)\sin\theta_k] + N(t), \qquad (3.30)$$

где $\tilde{u}(t)$ — преобразование Гильберта от $u(t)$: $\theta_k = \omega_0\tau$ — случайная начальная фаза. Распределение вероятностей θ_k предполагается заданным, чаще всего его задают равномерным на интервале от 0 до 2π. Эта модель удовлетворительно описывает те же каналы, что и предыдущая, если фаза сигнала в них флуктуирует. Такая флуктуация вызывается небольшими изменениями протяженности канала, свойств среды, в которой проходит сигнал, а также фазовой нестабильностью опорных генераторов.

Одполучевой гауссовский канал с общими замираниями (флуктуациями амплитуд и фаз сигнала) также описывается формулой (3.30), но множитель K, как и фаза 0_k, считаются случайными процессами. Иными словами, случайными будут квадратурные компоненты

$$X = K\cos\theta_k;\ Y = K\sin\theta_k. \qquad (3.31)$$

При изменении квадратурных компонент $X(t)$, $Y(t)$ во времени принимаемое колебание

$$Z(t) = X(t)u(t) + Y(t)\tilde{u}(t) + N(t) = K(t)[u(t)\cos\theta_k(t) + \tilde{u}(t)\sin\theta_k(t)] + N(t). \qquad (3.32)$$

Как отмечалось на с. 94, одномерное распределение коэффициента передачи $K(t)$ может быть рэлеевским (3.25) или обобщенным рэлеевским (3.26). Такие каналы называют соответственно каналами с *рэлеевскими* или с *обобщенными рэлеевскими замираниями*. В более общем случае $K(t)$ имеет четырехпараметрическое распределение [8]. Такую модель называют обобщенной гауссовской. Модель однолучевого канала с замираниями достаточно хорошо описывает многие каналы радиосвязи в различных диапазонах волн, а также некоторые другие каналы.

Линейный канал со случайной передаточной функцией и гауссовским шумом представляет собой дальнейшее обобщение. В таком канале выходное колебание $Z(t)$ выражается через входной сигнал $U(t)$ и случайную импульсную реакцию канала $G(t, \tau)$:

$$Z(t) = \int_0^t G(t, \tau)U(t-\tau)\,d\tau + N(t). \qquad (3.33)$$

Эта модель достаточно универсальна как для проводной, так и для радиосвязи и описывает каналы с рассеянием во времени по частоте. Часто рассеянию во времени канала можно приписать дискретный характер (модель многолучевого канала) и вместо (3.33) пользоваться представлением

$$z(t)=\sum_{l=1}^{L}[X_l(t)\,u(t-\tau_l)+Y_l(t)\,\tilde{u}(t-\tau_l)]+N(t), \qquad (3.34)$$

где L — число лучей в канале; $X_l(t)$; $Y_l(t)$ — квадратурные компоненты передаточной функции канала для l-го луча, которые в пределах спектра узкополосного сигнала практически не зависят от ω.

Канал с рассеянием во времени и по частоте задан полностью, если помимо корреляционной функции шума $N(t)$ задана статистика случайной импульсной реакции канала $G(t, \tau)$ (или передаточной функции $K(t, i\omega)$) или статистика квадратурных компонент $X_l(t, \omega)$, $Y_l(t, \omega)$ по всем лучам. В зависимости от значений входящих сюда параметров в таком канале могут наблюдаться селективные замирания и эхо-сигналы.

Каналы со сложной аддитивной помехой (флуктуационной, сосредоточенной, импульсной) описываются любой из предыдущих моделей с добавлением дополнительных компонент аддитивной помехи. Их полное описание требует задания вероятностных характеристик всех компонент аддитивного шума, а также параметров канала. Эти модели наиболее полно отображают реальные каналы связи, однако редко используются в анализе ввиду их сложности.

Переходя к моделям дискретного канала, полезно напомнить, что в нем всегда содержится непрерывный канал, а также модем. Последний можно рассматривать как устройство, преобразующее непрерывный канал в дискретный. Поэтому, в принципе, можно вывести математическую модель дискретного канала из моделей непрерывного канала и модема. Такой подход часто является плодотворным, однако он приводит к довольно сложным моделям.

Рассмотрим простые модели дискретного канала, при построении которых свойства непрерывного канала и модема не учитывались. Следует, однако, помнить, что при проектировании системы связи имеется возможность варьировать в довольно широких пределах модель дискретного канала при заданной модели непрерывного канала путем изменения модема.

Модель дискретного канала содержит задание множества возможных сигналов на его входе и распределение условных вероятностей выходного сигнала при заданном входном. Здесь входным и выходным сигналами являются последовательности кодовых символов. Поэтому для определения возможных входных сигналов достаточно указать число m различных символов (основание кода), а также длительность T передачи каждого символа. Будем считать, что значение T одинаково для всех символов, что выполняется в большинстве современных каналов. Величина $v=1/T$ определяет количество символов, передаваемых в единицу времени. Как указывалось в § 1.5, она называется технической скоростью и измеряется в бодах. Каждый символ, поступивший на вход канала, вызывает появление одного символа на выходе, так что техническая скорость на входе и выходе канала одинакова[1].

---

[1] В реальных каналах это не всегда выполняется, так как при нарушении тактовой синхронизации модема число символов на выходе канала может оказаться больше или меньше, чем на входе. В данном курсе это обстоятельство не учитывается и синхронизация считается идеальной. Методы обеспечения синхронизации изучаются в специальных курсах.

В общем случае для любого n должна быть указана вероятность того, что при подаче на вход канала любой заданной последовательности $\mathbf{b}^{[n]}$ кодовых символов на выходе появится некоторая реализация случайной последовательности $\hat{\mathbf{B}}^{[n]}$. Кодовые символы обозначим числами от 0 до $m-1$, что позволит производить над ними арифметические операции. При этом все n-последовательности (векторы), количество которых равно m^n, образуют m^n-мерное конечное векторное пространство, если «сложение» понимать как поразрядное суммирование по модулю m и аналогично определить умножение на скаляр (целое число). Для частного случая $m = 2$ такое пространство было рассмотрено в § 2.6.

Введем еще одно полезное определение. Будем называть *вектором ошибки* поразрядную разность (разумеется, по модулю m) между принятым и переданным векторами. Это значит, что прохождение дискретного сигнала через канал можно рассматривать как сложение входного вектора с вектором ошибки. Вектор ошибки играет в дискретном канале примерно ту же роль, что и помеха в непрерывном канале. Таким образом, для любой модели дискретного канала можно записать, пользуясь сложением в векторном пространстве (поразрядным, по модулю m):

$$\hat{\mathbf{B}}^{[n]} = \mathbf{B}^{[n]} + \mathbf{E}^{[n]}, \tag{3.35}$$

где $\mathbf{B}^{[n]}$ и $\hat{\mathbf{B}}^{[n]}$ — случайные последовательности из n символов на входе и выходе канала; $\mathbf{E}^{[n]}$ — случайный вектор ошибки, который в общем случае зависит от $\mathbf{B}^{[n]}$. Различные модели отличаются распределением вероятностей вектора $\mathbf{E}^{[n]}$. Смысл вектора ошибки особенно прост в случае двоичных каналов ($m=2$), когда его компоненты принимают значения 0 и 1. Всякая единица в векторе ошибок означает, что в соответствующем месте передаваемой последовательности символ принят ошибочно, а всякий нуль означает безошибочный прием символа. Количество ненулевых символов в векторе ошибок называется его *весом*. Образно говоря модем, осуществляющий переход от непрерывного канала к дискретному, преобразует помехи и искажения непрерывного канала в поток ошибок.

Перечислим наиболее важные и достаточно простые модели дискретных каналов.

Симметричный канал без памяти определяется как дискретный канал, в котором каждый переданный кодовый символ может быть принят ошибочно с фиксированной вероятностью p и правильно с вероятностью $1-p$, причем в случае ошибки вместо переданного символа b может быть с равной вероятностью принят любой другой символ. Таким образом, вероятность того, что принят символ $\hat{b}_j$, если был передан b_i, равна

$$p(\hat{b}_j \mid b_i) = \begin{cases} p/(m-1) & \text{при } i \neq j; \\ 1-p & \text{при } i = j. \end{cases} \tag{3.36}$$

Термин «без памяти» означает, что вероятность ошибочного приема символа не зависит от предыстории, т. е. от того, какие символы передавались до него и как они были приняты. В дальнейшем, для сокращения, вместо «вероятность ошибочного приема символа» будем говорить «вероятность ошибки».

Очевидно, что вероятность любого n-мерного вектора ошибки в таком канале

$$p\left(\mathbf{E}^{[n]}\right) = [p/(m-1)]^l (1-p)^{n-l}, \tag{3.37}$$

где l — количество ненулевых символов в векторе ошибки (вес вектора ошибки)[1]. Вероятность того, что произошло l каких угодно ошибок, расположенных как угодно на протяжении последовательности длины n, определяется формулой Бернулли

$$p(l) = C_n^l \left(\frac{p}{m-1}\right)^l (1-p)^{n-l}, \tag{3.38}$$

где $C_n^l = \dfrac{n!}{l!\,(n-l)!}$ — биномиальный коэффициент, равный числу различных сочетаний l ошибок в блоке длиной n.

Эту модель называют также *биномиальным каналом*. Она удовлетворительно описывает канал, возникающий при определенном выборе модема, если в непрерывном канале отсутствуют замирания, а аддитивный шум белый (или, по крайней мере, квазибелый). Вероятности переходов в двоичном симметричном канале схематически показаны в виде графа на рис. 3.3.

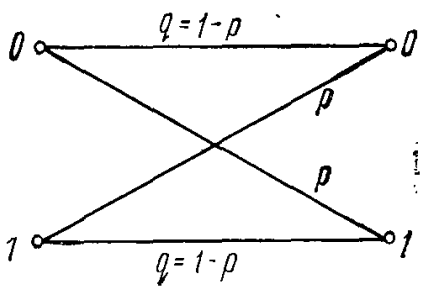


Рис. 3.3. Переходные вероятности в двоичном симметричном канале

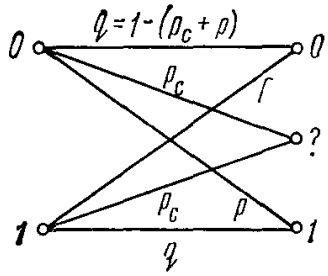


Рис. 3.4. Переходные вероятности в двоичном симметричном канале со стиранием

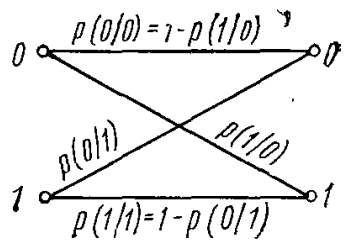


Рис. 3.5. Переходные вероятности в двоичном несимметричном канале

Симметричный канал без памяти со стиранием отличается от предыдущего тем, что алфавит на выходе канала содержит дополнительный $(m+1)$-й символ, обозначаемый знаком «?». Этот символ появляется тогда, когда 1-я решающая схема (демодулятор) не может надежно опознать переданный символ. Вероятность такого *отказа от решения* или *стирания* символа p_c в данной модели постоянна и не зависит от передаваемого

[1] В двоичном канале вес вектора совпадает с его нормой, определенной формулой (2.104).

символа. За счет введения стирания удается значительно снизить вероятность ошибки, иногда ее даже считают равной нулю. На рис. 3.4 схематически показаны вероятности переходов в такой модели.

Несимметричный канал без памяти характеризуется, как и предыдущие модели, тем, что ошибки возникают в нем независимо друг от друга, однако вероятности ошибок зависят от того, какой символ передается. Так, в двоичном несимметричном канале вероятность $p(1|0)$ приема символа «1» при передаче символа «0» не равна вероятности $p(0|1)$ приема «0» при передаче «1» (рис. 3.5). В этой модели вероятность вектора ошибки зависит от того, какая последовательность символов передается.

Марковский канал представляет собой простейшую модель дискретного канала с памятью. В ней вероятность ошибки образует простую цепь Маркова, т. е. зависит от того, правильно или ошибочно принят предыдущий символ, но не зависит от того, какой символ передается.

Такой канал, например, возникает, если в непрерывном канале с гауссовским шумом (с определенной или неопределенной фазой) используется относительная фазовая модуляция (см. ниже, § 4.5).

Канал с аддитивным дискретным шумом является обобщением моделей симметричных каналов. В такой модели вероятность вектора ошибки $\mathbf{E}^{[n]}$ не зависит от передаваемой последовательности. Вероятность каждого вектора ошибки считается заданной и, вообще говоря, не определяется его весом. Во многих каналах из двух векторов с одинаковым весом более вероятным оказывается такой, в котором единицы расположены близко друг к другу, т. е. имеется тенденция к группированию ошибок.

Частным случаем такого канала является канал с переменным параметром (КПП). В этой модели вероятность ошибки для каждого символа является функцией некоторого параметра $\mu(t)$, представляющего случайную последовательность, дискретную или непрерывную, с известными распределениями вероятностей, в частности с известной корреляционной функцией. Параметр μ может быть скалярным или векторным. Можно сказать, что μ определяет состояние канала. Такая модель имеет много разновидностей. Одной из них является модель Гильберта, в которой μ принимает лишь два значения — μ_1 и μ_2, а вероятность ошибки при $\mu=\mu_1$ равна нулю, а при $\mu=\mu_2$ равна 0,5. Заданы вероятности переходов из состояния μ_1 в μ_2 и наоборот. В таком канале все ошибки происходят при $\mu=\mu_2$ и поэтому очень тесно группируются. Существуют и более сложные модели КПП, например модель Попова — Турина. Они изучаются в специальных курсах. Память в КПП определяется интервалом корреляции параметра μ.

Канал с неаддитивным шумом и с памятью. Канал с межсимвольной интерференцией. Вероятность ошибки в нем зависит от передаваемых символов, как и в модели несимметричного канала без памяти, но не от того (или не только от того) символа, для которого определяется вероятность ошибки, а от символов, которые передавались до него.

3.5. УРАВНЕНИЯ СОСТОЯНИЯ. МАРКОВСКИЕ МОДЕЛИ КАНАЛОВ СВЯЗИ

Рассмотрим другой подход к построению математических моделей каналов. Выше соотношения между входным и выходным сигналами задавались интегральными преобразованиями (например, интегралом Дюамеля). При этом для нахождения выходного сигнала требуется знать помимо характеристик цепи (канала) также входной сигнал, действовавший на нее на *всем промежутке* его существования, до текущего момента t.

Во многих случаях более гибким является такое описание, при котором вся предыстория до некоторого фиксированного момента времени t_0 заменяется заданием некоторого начального состояния цепи. Зная характеристики цепи, начальное состояние и сигнал, действующий только на промежутке от t_0 до t, можно последовательно определить как сигнал на выходе, так и новое состояние цепи в любой момент времени $t > t_0$.

Читатели знакомы с подобным подходом из теории дифференциальных уравнений, в которой искомая функция определяется как самим уравнением, так и определенными начальными условиями, число которых равно порядку уравнения. Излагаемый в данном параграфе *метод переменных состояния* иллюстрируется на примерах цепей, описываемых с помощью линейных дифференциальных уравнений.

Множество величин, однозначно определяющих поведение цепи в некоторый момент t, содержащее минимальное число элементов n, называют *состоянием*, а сами элементы этого множества — переменными состояния. Каждую из этих переменных обычно рассматривают как составляющую n-мерного *вектора состояний*. Для любой заданной цепи можно составить два уравнения, позволяющих по состоянию в момент t_0 и сигналу, поступающему на вход, найти состояние в момент $t > t_0$ и выходной сигнал. Первое из них называется *уравнением состояния*, а второе — *уравнением наблюдения*.

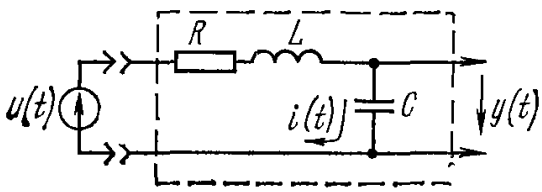


Рис. 3.6. Последовательный колебательный контур

Для иллюстрации основных положений метода переменных состояния рассмотрим простой пример — линейную RLC-цепь (рис. 3.6), в которой выходное напряжение $y(t)$ связано с входным напряжением $u(t)$ дифференциальным уравнением

$$d^2 y(t)/dt^2 + 2\alpha\, dy(t)/dt + \omega_0^2 y(t) = \omega_0^2 u(t), \qquad (3.39)$$

где $2\alpha = R/L$; $\omega_0 = 1/(LC)$. Ток в цепи

$$i(t) = C\, dy(t)/dt. \qquad (3.40)$$

Состояние этой цепи в любой момент времени t_0 характеризуется двумя параметрами: $i(t_0)$ — током, протекающим через индуктивность L, и $y(t_0)$ — падением напряжения на емкости C. Значения $i(t_0)$, $y(t_0)$ содержат достаточную информацию о предыстории цепи, связанной с прошлыми воздействиями $u(t)$, $t \leqslant t_0$, которая необходима для определения будущих значений выходного процесса $y(t)$, $t > t_0$, при заданных воздействиях $u(t)$, $t \geqslant t_0$. Таким образом, $dy(t)/dt$ и $y(t)$ можно интерпретировать как переменные состояния, а дифференциальное уравнение (3.39) — как уравнение состояния, которое обычно приводят к форме векторного дифференциального уравнения первого порядка. При замене переменных

$$x_1(t) = y(t), \quad x_2(t) = dy(t)/dt \qquad (3.41),\ (3.42)$$

уравнение (3.39) эквивалентно системе дифференциальных уравнений первого порядка:

$$dx_1(t)/dt = x_2(t), \quad \frac{d}{dt} x_2(t) = -\omega_0^2 x_1(t) - 2\alpha x_2(t) + \omega_0^2 u(t), \qquad (3.43),\ (3.44)$$

которая с учетом правил векторно-матричных преобразований допускает компактное представление

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{F}\,\mathbf{x}(t) + \mathbf{G}\,u(t), \tag{3.45}$$

где

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}; \quad \mathbf{F} = \begin{pmatrix} 0 & 1 \\ -\omega_0^2 & -2\alpha \end{pmatrix}; \quad \mathbf{G} = \begin{pmatrix} 0 \\ \omega_0^2 \end{pmatrix}. \tag{3.46}$$

При этом уравненне наблюдения имеет вид (3.41), или в векторной форме

$$y(t) = \mathbf{H}\mathbf{x}(t) = (1,\ 0)\,\mathbf{x}(t). \tag{3.47}$$

Заметим, что выбранный нами вектор состояния $\mathbf{x}(t)$ не является единственно возможиым. Любое несингулярное линейное преобразование вектора $\mathbf{x}(t)$ приводит к другому вектору состояний.

Важной особенностью метода переменных состояний является возможность непосредственного моделирования систем, описываемых уравнениями состояния, с помощью аналогового или цифрового вычислительного устройства. На рис. 3.7 показаиа модель системы уравнений (3.43) — (3.44). При построеннн такой

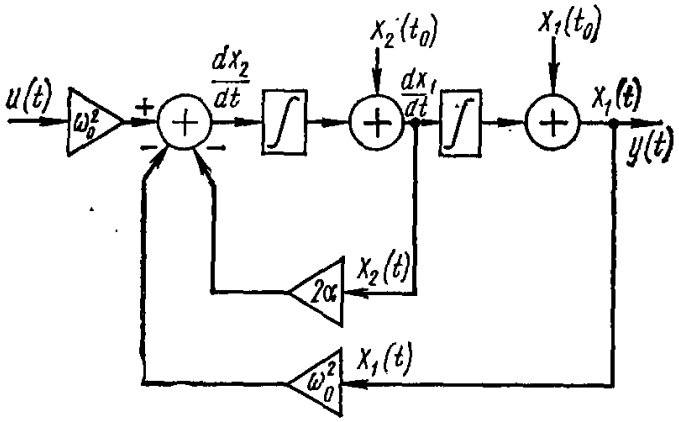


Рис. 3.7. Моделирование уравнений состояния линейной системы 2-го порядка (последовательного колебательного контура)

схемы удобно рассуждать следующим образом. Пусть в некоторых точках присутствуют входной сигнал $u(t)$ и переменные состояния $x_1(t)$ и $x_2(t)$. Соеднним эти точкн сумматорами, усилителямн и интеграторами так, чтобы соотношения между ними соответствовалн ур-ниям (3.43), (3.44). Из первого уравнення следует, что, подав на вход интегратора $x_2(t)$, мы должны получить, с точностью до постоянной, $x_1(t)$. Эта постоянная определяется начальным условием и равна $x_1(t_0)$. Затем осуществляем операции, запнсанные в правой части второго уравнения, умиожив $u(t)$ н $x_1(t)$ на ω^2_0, а $x_2(t)$ на 2α (с помощью усилителей с соответствующнми коэффициентами усиления) и сложив полученные результаты с учетом знаков. Проинтегрировав полученную сумму и прнбавнв к ней постоянную $x_2(t_0)$ (начальное условне), получнм $x_2(t)$. Таким образом, все точки в схеме соединились в соответствии с уравиениями (3.43) н (3.44).

Еслн на такую схему-модель подать входной снгнал $u(t)$, то на выходе получится выходной сигнал $y(t)$. Однако это не представляет большого интереса, поскольку то же самое можно сделать без моделирования, исследуя экспериментально исходную цепь (в данном случае рис. 3.6). Значительно важнее то, что с помощью модели можно решить обратную задачу — по выходному (наблюдаемому) сигналу найти входной.

В более общем случае аналогнчные матрнчные уравнения в форме (3.45) и (3.47) можно построить для цепей более высокого порядка, в том числе нелн-

нейных и с переменными параметрами. Отличие будет лишь в размерности матриц и в том, что они могут быть функциями времени (для цепей с переменными параметрами) и состояния (для нелинейных цепей). Если на цепь воздействует несколько входных и несколько выходных сигналов, то их также рассматривают как компоненты вектор-функций $\mathbf{u}(t)$ и $\mathbf{y}(t)$. В самом общем случае уравнения состояния и наблюдения принимают в векторной форме следующий вид:

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(t, \mathbf{x})\cdot\mathbf{x}(t) + \mathbf{G}(t, \mathbf{x})\cdot\mathbf{u}(t); \tag{3.48}$$

$$\mathbf{y}(t) = \mathbf{H}(t, \mathbf{x})\cdot\mathbf{x}(t), \quad t \geqslant t_0; \quad \mathbf{x}(t_0) = \mathbf{x}_0. \tag{3.49}$$

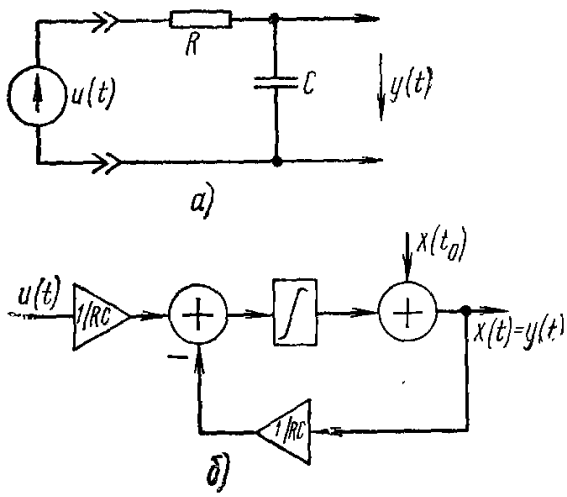


Рис. 3.8. Цепь *RC* (*а*); модель уравнений состояния цепи *RC* (*б*)

Каждое из матричных уравнений представляет, в сущности, систему дифференциальных уравнений, число которых для уравнений состояния равно количеству переменных состояния (порядку цепи), а для уравнений наблюдения — количеству выходов цепи [1].

Отметим еще простейший случай линейной цепи 1-го порядка, например интегрирующей *RC*-цепи (рис. 3.8*а*). В этом случае уравнения состояния (3.45) и наблюдения (3.47) оказываются скалярными, а в качестве единственной переменной состояния $x(t)$ удобно принять напряжение на конденсаторе, которое совпадает с выходным сигналом $y(t)$. Уравнения имеют вид

$$dx/dt = -(1/RC)\,x(t) + (1/RC)\,u(t); \quad y(t) = x(t), \tag{3.50}$$

что совпадает с (3.45) и (3.47) при $\mathbf{F} = -(1/RC)$, $\mathbf{G} = (1/RC)$, $\mathbf{H} = 1$. Соответствующая модель дана на рис. 3.8*б*.

Одно из приложений метода переменных состояния связано с возможностью конструктивного описания случайных процессов. Оно состоит в том, что случайный процесс $\mathbf{Y}(t)$ с заданными вероятностными характеристиками представляется как выход некоторой динамической системы, возбуждаемой другим случайным процессом с более простой вероятностной структурой $\mathbf{U}(t)$. Обычно в качестве порождающего используется гауссовский процесс $\mathbf{U}(t)$ типа белого шума с нулевым средним и корреляционной функцией

$$\mathrm{B}_{\mathbf{U}}(t, t+\tau) = \mathrm{M}\{\mathbf{U}(t)\,\mathbf{U}'(t+\tau)\} = \mathbf{Q}\,\delta(\tau), \tag{3.51}$$

где $\mathbf{Q}$ — симметричная, неотрицательно определенная матрица, а штрих обозначает транспонирование. Можно показать, что при воздействии случайного процесса $\mathbf{U}(t)$, удовлетворяющего (3.51), на схему, описываемую (3.48) и (3.49), где функции $\mathbf{F}(t)$, $\mathbf{H}(t)$ и $\mathbf{G}(t)$ удовлетворяют условиям непрерывности и ограниченности, процессы $\mathbf{X}(t)$, а также $\{\mathbf{X}(t), \mathbf{Y}(t)\}$ являются марковскими, переходные плотности вероятностей которых $w(\mathrm{x}, t|\mathrm{x}_0, t_0)$ и $w(\mathrm{x}, \mathrm{y}, t|\mathrm{x}_0, \mathrm{y}_0, t_0)$ подчинены соответствующим дифференциальным уравнениям в частных производных Колмогорова — Фокера — Планка (2.30) (см. § 2.1). Если гауссовский порождающий процесс $\mathbf{U}(t)$ воздействует на линейную цепь, то и выходной процесс $\mathbf{Y}(t)$ будет гауссовским. Он будет также стационарным, если формирующая система является линейной с постоянными параметрами. Распределение вероят-

[1] Описание систем с дискретным временем в основном аналогично описанию систем с непрерывным временем; при этом дифференциальное уравнение состояний сводится к уравнению в конечных разностях.

ностей процесса $\mathbf{Y}(t)$ будет негауссовским, если он сформирован нелинейной системой (3.48) — (3.49).

В частности, если на вход цепи рис. 3.8 поступает нормальный белый шум, то выходным процессом будет марковский гауссовский процесс с нормированной корреляционной функцией (2.31).

Метод переменных состояния с успехом применяется для описания стохастических цепей (каналов) со случайно изменяющимися параметрами. Для этого некоторые элементы системных функций (матриц) **F**, **G**, **H** следует рассматривать как случайные функции. Кроме того, в уравнении наблюдения следует учитывать компоненту аддитивных помех.

Метод переменных состояний дает универсальный подход для моделирования (в рамках весьма широкой марковской модели) каналов передачи информации (систем связи) для самых различных сообщений, способов кодирования и модуляции (как линейной, так и нелинейной), линий связи с детерминированными и случайными параметрами, рассеянием сигналов, аддитивными шумами (как гауссовскими, так и негауссовскими). Более существенно то обстоятельство, что, представляя наблюдаемые (анализируемые в месте приема) случайные марковские процессы с помощью дифференциальных уравнений (уравнений состояния) можно решить обратную задачу, т. е. получить дифференциальные уравнения для оценки сообщений, заключенных в этих процессах. Такие оценки с помощью аналоговой или цифровой техники значительно проще, чем оценки, следующие из интегральных уравнений. Это будет показано в § 6.2 и 6.10 при рассмотрении фильтрации сигналов на фоне шумов.

3.6. ПРОПУСКНАЯ СПОСОБНОСТЬ КАНАЛА СВЯЗИ

В любой системе связи через канал передается информация. Скорость передачи информации была определена в § 2.9. Эта скорость зависит не только от самого канала, но и от свойств подаваемого на его вход сигнала и поэтому не может характеризовать канал как средство передачи информации. Попытаемся найти способ оценки способности канала передавать информацию. Рассмотрим вначале дискретный канал, через который передаются в единицу времени v символов из алфавита объемом m. При передаче каждого символа в среднем по каналу проходит следующее количество информации [см. (2.135) и (2.140)]:

$$I(B, \hat{B}) = H(B) - H(B \mid \hat{B}) = H(\hat{B}) - H(\hat{B} \mid B), \qquad (3.52)$$

где B и $\hat{B}$ — случайные символы на входе и выходе канала. Из четырех фигурирующих здесь энтропий $H(B)$ — собственная информация передаваемого символа — определяется источником дискретного сигнала[1] и не зависит от свойств канала. Остальные три энтропии в общем случае зависят как от источника сигнала, так и от канала.

Представим себе, что на вход канала можно подавать символы от разных источников, характеризуемых различными распределениями вероятностей $P(B)$ (но, конечно, при тех же значениях m и v). Для каждого такого источника количество информации, переданной по каналу, принимает свое значение. Максимальное количество переданной информации, взятое по всевозможным источ-

[1] Источником дискретного сигнала в системе связи (см. рис. 1.5) является совокупность источника сообщения и кодера.

никам входного сигнала, характеризует сам канал и называется *пропускной способностью канала.* В расчете на один символ

$$C_{\text{символ}} = \max_{P(B)} I(B, \hat{B}), \text{ бит/символ,} \tag{3.53}$$

где максимизация[1] производится по всем многомерным распределениям вероятностей $P(B)$. Можно также определить пропускную способность C канала в расчете на единицу времени (секунду):

$$C = \max_{P(B)} I'(B, \hat{B}) = vC_{\text{символ}}, \text{ бит/с.} \tag{3.54}$$

Последнее равенство следует из аддитивности энтропии. В дальнейшем везде, где это особо не оговорено, будем под пропускной способностью понимать пропускную способность в расчете на секунду.

В качестве примера вычислим пропускную способность симметричного канала без памяти, для которого переходные вероятности заданы формулой (3.36). Согласно (3.52) и (3.53)

$$C_{\text{симв}} = \max_{P(B)} [H(\hat{B}) - H(\hat{B} \mid B)]. \tag{3.55}$$

Величина $H(\hat{B} \mid B) = \mathrm{M}\left\{\log \dfrac{1}{P(\hat{b}_j \mid b_i)}\right\}$ в данном случае легко вычисляется, поскольку условная переходная вероятность $P(\hat{b}_j \mid b_i)$ принимает только два значения: $p/(m-1)$, если b_i, и $1-p$, если $\hat{b}_j = b_i$. Первое из этих значений возникает с вероятностью p, а второе с вероятностью $1-p$. К тому же, поскольку рассматривается канал без памяти, результаты приема отдельных символов независимы друг от друга. Поэтому

$$H(\hat{B} \mid B) = \mathrm{M}\left\{\log \frac{1}{P(\hat{b} \mid b)}\right\} = p \log \frac{m-1}{p} + (1-p) \log \frac{1}{1-p}. \tag{3.56}$$

Следовательно, $H(\hat{B} \mid B)$ не зависит от распределения вероятности B, а определяется только переходнымн вероятностями канала. Это свойство сохраняется для всех моделей канала с аддитивным шумом.

Подставив (3.56) в (3.55), получнм

$$C_{\text{симв}} = \max_{P(B)} \left[H(\hat{B}) - p \log \frac{m-1}{p} - (1-p) \log \frac{1}{1-p}\right].$$

Поскольку в правой части только член $H(\hat{B})$ зависнт от распределения вероятностей $P(B)$, то максимизировать необходимо его. Максимальное значение $H(\hat{B})$ согласно (2.123) равно $\log m$ и реализуется оно тогда, когда все принятые символы $\hat{b}_j$ равновероятны н независимы друг от друга. Легко убедиться, что это условие удовлетворяется, если входные снмволы равновероятны и независимы, поскольку

$$P(b_i) = \frac{1}{m} \text{ и } P(\hat{b}_j) = \sum_{i=1} P(b_i)\, P(\hat{b}_j / b_i) =$$

[1] Если такого максимума не существует (что может быть при бесконечном чнсле возможных источников), то пропускная способность определяется как наименьшая верхняя грань $\sup I(B, \hat{B})$, т. е. такая величиина, к которой $I(B, \hat{B})$ может сколь угодно приблизиться, но не может ее превзойти.

$$= \frac{1}{m} \sum_{i=1}^{m} P(\hat{b}_j / b_i) = \frac{1}{m}\left[(1-p) + (m-1)\frac{p}{m-1}\right] = \frac{1}{m}.$$

При этом $H(B) = \log m$ и

$$C_{\text{симв}} = \log m + p \log \frac{p}{m-1} + (1-p)\log(1-p). \qquad (3.57)$$

Отсюда пропускная способность в расчете на секунду

$$C = v\left[\log m + p \log \frac{p}{m-1} + (1-p)\log(1-p)\right]. \qquad (3.58)$$

Для двоичного симметричного канала ($m=2$) пропускная способность в двоичных единицах в секунду

$$C = v\,[1 + p \log p + (1-p)\log(1-p)]. \qquad (3.59)$$

Зависимость C/v от p согласно (3.59) показана на рис. 3.9.

При $p=1/2$ пропускная способность двоичного канала $C=0$, поскольку при такой вероятности ошибки последовательность выходных двоичных символов можно получить, совсем не передавая сигналы по каналу, а выбирая их наугад (например, по результатам бросания монеты), т. е. при $p=1/2$ последовательности на выходе и входе канала независимы. Случай $C=0$ называют обрывом канала. То, что пропускная способность при $p=1$ в двоичном канале такая же, как при $p=0$ (канал без шумов), объясняется тем, что при $p=1$ достаточно все выходные символы инвертировать (т. е. заменить 0 на 1 и 1 на 0), чтобы правильно восстановить входной сигнал.

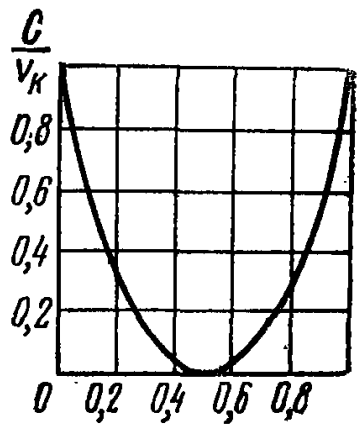


Рис. 3.9. Зависимость пропускной способности двоичного симметричного каиала без памяти от вероятности ошибочного прнема символа

Пропускная способность непрерывного канала вычисляется аналогично. Пусть, например, канал имеет ограниченную полосу пропускания шириной F. Тогда сигналы $U(t)$ и $Z(t)$ на входе и выходе канала по теореме Котельникова определяются своими отсчетами, взятыми через интервал $1/(2F)$, и поэтому информация, проходящая по каналу за некоторое время T, равна сумме количеств информации, переданных за каждый такой отсчет[1]. Пропускная способность канала на один такой отсчет

$$C_{\text{отсч}} = \max_{w(u)} I(U, Z) = \max_{w(u)} [h(U) - h(U \mid Z)] = \max_{w(u)} [h(Z) - h(Z \mid U)]. \qquad (3.60)$$

[1] Можно вместо ряда Котельникова использовать разложенне сигналов по любому ортогональному базису и рассмотреть количество передаваемой иифорлации на каждый член ряда. Однако удобиее применять канонические разложения, так как при этом коэффициенты не коррелированы.

Здесь U и Z — случайные величины — сечения процессов $U(t)$ и $Z(t)$ на входе и выходе канала и максимум берется по всем допустимым входным сигналам, т. е. по всем распределениям u.

Пропускная способность C определяется как сумма значений $C_{\text{отсч}}$, взятая по всем отсчетам за секунду. При этом, разумеется, дифференциальные энтропии в (3.60) должны вычисляться с учетом вероятностных связей между отсчетами.

Вычислим, например, пропускную способность непрерывного канала без памяти с аддитивным белым гауссовским шумом, имеющим полосу пропускания шириной F, если средняя мощность сигнала (дисперсия U) не превышает заданной величины $P_{\text{с}}$. Мощность (дисперсию) шума в полосе F обозначим $P_{\text{ш}}$. Отсчеты входного и выходного сигналов, а также шума N связаны равенством

$$Z = U + N, \tag{3.61}$$

и так как N имеет нормальное распределение с нулевым математическим ожиданием, то и условная плотность вероятности $w(z/u)$ при фиксированном u будет также нормальной — с математическим ожиданием u и дисперсией $P_{\text{ш}}$.

Найдем пропускную способность на один отсчет:

$$C_{\text{отсч}} = \max_{w(u)} [h(Z) - h(Z \mid U)].$$

Согласно (2.152) дифференциальная энтропия $h(Z|U)$ нормального распределения $w(z|u)$ не зависит от математического ожидания и равна $\log \sqrt{2\pi e P_{\text{ш}}}$. Поэтому для нахождения $C_{\text{отсч}} = \max\limits_{w(u)} [h(z) - \log \sqrt{2\pi e P_{\text{ш}}}]$ нужно найти такую плотность распределения $w(u)$, при которой максимизируется $h(Z)$. Из (3.61), учитывая, что U и N — независимые случайные величины, имеем

$$\mathrm{D}(Z) = \mathrm{D}(U) + \mathrm{D}(N) = P_{\text{с}} + P_{\text{ш}}. \tag{3.62}$$

Таким образом, дисперсия Z фиксирована, так как $P_{\text{с}}$ и $P_{\text{ш}}$ заданы. Согласно (2.153), при фиксированной дисперсии максимальная дифференциальная энтропия обеспечивается нормальным распределением. Из (3.61) видно, что при нормальном одномерном распределении U распределение Z будет также нормальным и, следовательно,

$$\max_{w(u)} h(Z) = \log \sqrt{2\pi e (P_{\text{с}} + P_{\text{ш}})},$$

откуда

$$C_{\text{отсч}} = \log \sqrt{2\pi e (P_{\text{с}} + P_{\text{ш}})} - \log \sqrt{2\pi e P_{\text{ш}}} = \frac{1}{2} \log \frac{P_{\text{с}} + P_{\text{ш}}}{P_{\text{ш}}}. \tag{3.63}$$

Переходя к пропускной способности C в расчете на секунду, заметим, что информация, переданная за несколько отсчетов, максимальна в том случае, когда отсчеты сигналов независимы. Этого можно достичь, если сигнал $U(t)$ выбрать так, чтобы его спектральная плотность была равномерной в полосе F. Как было показано в § 2.2 [ф-ла (2.48)], отсчеты, разделенные интервалами, кратными $1/(2F)$, взаимно некоррелированны, а для гауссовских величин некоррелированность означает независимость.

Поэтому пропускную способность C (за секунду) можно найти, сложив пропускные способности (3.63) для $2F$ независимых отсчетов:

$$C = 2FC_{\text{отсч}} = F \log (1 + P_{\text{с}}/P_{\text{ш}}). \tag{3.64}$$

Она реализуется, если $U(t)$ — гауссовский процесс с равномерной спектральной плотностью в полосе частот F (квазибелый шум).

Из формулы (3.64) видно, что если бы мощность сигнала P_c не была ограничена, то пропускная способность была бы бесконечной. Пропускная способность равна нулю, если отношение сигнал/шум $P_c/P_ш$ в канале равно нулю. С ростом этого отношения пропускная способность увеличивается неограниченно, однако медленно, вследствие логарифмической зависимости.

Соотношение (3.64) часто называют формулой Шеннона. Эта формула имеет важное значение в теории информации, так как определяет зависимость пропускной способности рассматриваемого непрерывного канала от таких его технических характеристик, как ширина полосы пропускания и отношение сигнал/шум. Формула Шеннона указывает на возможность обмена полосы пропускания на мощность сигнала и наоборот. Однако поскольку C зависит от F линейно, а от $P_c/P_ш$ — по логарифмическому закону, компенсировать возможное сокращение полосы пропускания увеличением мощности сигнала, как правило, нецелесообразно. Более эффективным является обратный обмен мощности сигнала на полосу пропускания.

Выясним, как меняется пропускная способность гауссовского канала с изменением полосы пропускания. Для этого выразим мощность шума в канале через его одностороннюю спектральную плотность N_0. Имеем $P_ш = N_0F$; поэтому формулу (3.64) можно представить в виде

$$C = F\log\left(1+\frac{P_c}{N_0F}\right) = F\log e\ln\left(1+\frac{P_c}{N_0F}\right). \qquad (3.65)$$

При увеличении F пропускная способность C сначала быстро возрастает, а затем асимптотически стремится к пределу:

$$C_\infty = \lim_{F\to\infty} C = (P_c/N_0)\log e, \text{ бит/с}. \qquad (3.66)$$

Результат (3.66) получается очень просто, если учесть, что при $|\varepsilon| \ll 1$ $\ln(1+\varepsilon)\approx\varepsilon$. Зависимость C от F показана на рис. 3.10. Как следует из формулы (3.66), для передачи заданного количества информации по каналу с шумами отношение энергии сигнала к спектральной плотности шума $h^2 = P_cT/N_0$ должно превышать некоторую пороговую величину. В самом деле, если на передачу сообщения затрачено время T, то среднее количество переданной информации $TI'(U, Z) < TC_\infty$, так как пропускная способность канала при любой полосе F не может превысить предельное значение (3.66). Таким образом,

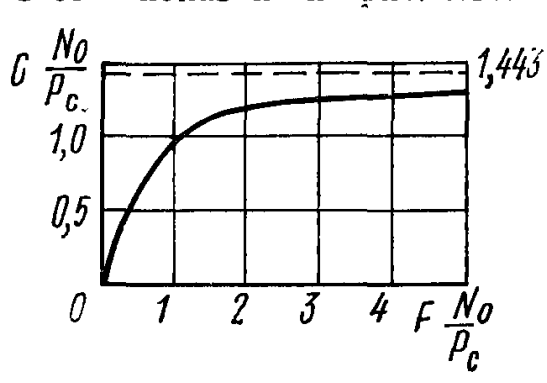


Рис. 3.10. Зависимость нормированной пропускной способности гауссовского канала от его полосы пропускания

$$TI'(U, Z) < (P_cT/N_0)\log e,$$

и, следовательно, для передачи 1 бита информации необходима энергия сигнала $P_c T > N_0/\log e = N_0 \ln 2 \approx 0{,}693\, N_0$, или

$$h^2 > 0{,}693. \tag{3.67}$$

Максимальный объем информации, которую можно в среднем передать по непрерывному каналу за время T_k, равен $V_k = T_k C$. Для гауссовского канала

$$V_k = T_k F_k \log(1 + P_c/P_ш). \tag{3.68}$$

Заметим, что при $P_c/P_ш \gg 1$ выражение (3.68) совпадает с характеристикой (1.2), названной в § 1.2 емкостью (объемом) канала.

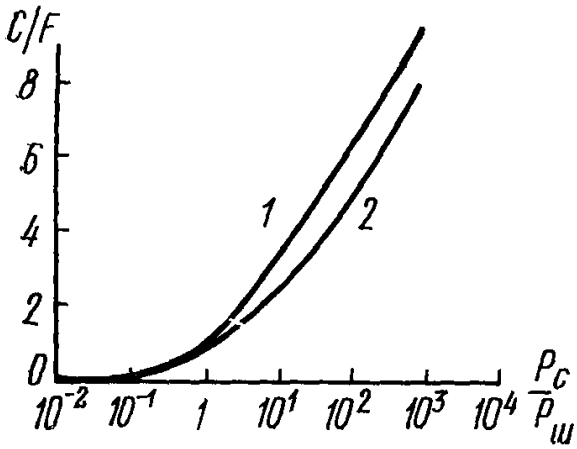


Рис. 3.11. Зависимость пропускной способности от отношения сигнал/шум для канала с постоянными параметрами (1) и с рэлеевскими замираниями (2)

Следует подчеркнуть, что формула Шеииоиа (3.64) справедлива только для каиала с постоянными параметрами и аддитивным гауссовским белым (или квазибелым) шумом. Если распределеиие аддитивной помехи не являетси нормальным или же ее спектр иеравиомереи в полосе пропускания канала, то его пропускная способиость больше, чем вычислениаи по формуле (3.64). Мультипликативные помехи (замирания сигнала) обычио снижают пропускную способность канала.

На рис. 3.11 показаны зависимости C/F от среднего отношеиия $P_c/P_ш$ для каиала с постоянными параметрами (*1*) и каиала с рэлеевскимн замираниями (*2*). Из анализа кривых рис. 3.11 следует, что медленные рэлеевские замирания уменьшают пропускную способность канала не более чем на 17%.

3.7. ТЕОРЕМА КОДИРОВАНИЯ ДЛЯ КАНАЛА С ПОМЕХАМИ

Пропускная способность канала, определенная в предыдущем параграфе, характеризует потенциальные возможности передачи информации. Они раскрываются в фундаментальной теореме теории информации, известной как *основная теорема кодирования* К. Шеннона. Применительно к дискретному источнику она формулируется так: *если производительность источника сообщений* $H'(A)$ *меньше пропускной способности канала* C:

$$H'(A) < C, \tag{3.69}$$

то существует способ кодирования (преобразования сообщения в сигнал на входе) и декодирования (преобразования сигнала в сообщение на выходе канала), при котором вероятность ошибочного декодирования и ненадежность $H(A|\bar{A})$ *могут быть сколь угодно малы. Если же*

$$H'(A) > C, \tag{3.70}$$

то таких способов не существует.

Таким образом, согласно теореме К. Шеннона конечная величина C — это предельное значение скорости безошибочной передачи информации по каналу. Эта теорема, к сожалению, неконструктивна, т. е. она не указывает конкретного способа кодирования, существование которого доказывается. Тем не менее, значение теоремы трудно переоценить, ибо она в корне изменила воззрения на принципиальные возможности техники связи. До К. Шеннона считалось, что в канале с шумами можно обеспечить сколь угодно малую вероятность ошибки только при неограниченном уменьшении скорости передачи информации. Таков, скажем, путь повышения верности связи за счет повторении символов в канале без памяти.

Например, сообщения двоичного источника $a_1=0$ и $a_2=1$ можно передавать по двоичному симметричному каналу с вероятностью ошибок $p<0{,}5$ двумя кодовыми комбинациями, содержащими соответственно n единиц или n нулей:

$$\mathbf{b}_1=\underbrace{0\ 0\ 0\ldots0}_{n},\quad \mathbf{b}_2=\underbrace{1\,1\,1\ldots1}_{n}. \tag{3.71}$$

Если в месте приема регистрировать 1 или 0 по большинству этих знаков в кодовой комбинации (мажоритарное декодирование), то ясно, что ошибка произойдет при условии, если в кодовой комбинации неверно будет принято $n/2$ или более символов. Согласно закону больших чисел вероятность уклонения числа ошибок m в кодовой комбинации длины n от их математического ожидании np стремится к нулю при $n\to\infty$:

$$\lim P\,(|\,m-np\,|>\varepsilon)=0- \tag{3.72}$$

при сколь угодно малом положительном ε.

Поскольку $np<n/2$, код (3.71) обеспечит при $n\to\infty$ безошибочный прием, однако одновременно при этом и скорость передачи информации по каиалу стремится к нулю, тогда как согласно теореме К. Шеннона существуют коды, которые, в отличие от кода (3.71), обеспечивают сколь угодно малую вероятность ошибки при конечиой скорости передачи информации.

Заметим, что условие (3.69) можно записать так:

$$v_{\text{и}}\,H\,(A)<v\,C_{\text{симв}}, \tag{3.73}$$

где $v_{\text{и}}$ — число сообщений, поступающих от источника в секунду.

Поэтому можно дать и другую формулировку основной теоремы К. Шеннона: *если среднее число символов канала на одно сообщение источника* $\bar{n}=v/v_{\text{и}}$ *удовлетворяет неравенству*

$$\bar{n}>H\,(A)/C_{\text{симв}}, \tag{3.74}$$

то существует способ кодирования и декодирования, обеспечивающий сколь угодно малую вероятность ошибки, если же

$$\bar{n}<H\,(A)/C_{\text{симв}},$$

такого способа нет. Как следует из (3.74) с учетом 3.57, при двоичном симметричном канале без памяти минимальное значение

$$\bar{n}_{\text{мин}}>\frac{H\,(A)}{1+p\log p+(1-p)\log(1-p)}.$$

Обсудим содержание теоремы Шеннона. Как отмечалось в § 2.9, для восстановления по пришедшему сигналу переданного сообщения необходимо, чтобы сигнал содержал о нем информацию, равную энтропии сообщения. Следовательно, для правильной передачи сообщения необходимо, чтобы скорость передачи информации была не меньше производительности источника. Так как по определению скорость передачи информации не превышает пропускной способности, то неравенство $H'(A)<C$ является необходимым условием для точной передачи сообщения. Но является ли это условие достаточным?

Конечно, при $C>H'(A)$ можно передавать такие сигналы, что $I'(B, \hat{B})$ достигнет значения $H'(A)$. Но $I'(B, \hat{B})$ — это скорость передачи информации *о сигнале* B, а не *о сообщении* A. Поэтому вопрос сводится к тому, можно ли установить такое соответствие (код) между сообщением A и сигналом B, чтобы вся информация, полученная на выходе канала о сигнале B, была в то же время и информацией о сообщении A?

Положительный ответ на этот вопрос очевиден только в тривиальном случае, когда в канале нет помех и сигнал B принимается безошибочно. При этом $I'(B, \hat{B})=H'(B)$, и если между A и B установлено взаимно однозначное соответствие, то по принятому сигналу можно однозначно восстановить сообщение. В общем же случае в канале имеются помехи и сигнал B принимается с ошибками, так что $I'(B, \hat{B})<H'(B)$. С учетом (3.54) и (3.69) отсюда следует, что $H'(B)>H'(A)$.

Это значит, что производительность сигнала B должна быть выше производительности источника сообщения A и, следовательно, B содержит кроме информации об A дополнительную собственную информацию. Часть информации о сигнале B в канале теряется. Наш вопрос сводится к следующему: можно ли осуществить кодирование так, чтобы терялась только дополнительная (избыточная) часть собственной информации B, а информация об A сохранялась?

Теорема Шеннона дает на этот вопрос почти положительный ответ, с той лишь поправкой, что скорость «утечки информации» (или ненадежность) $H'(A|\hat{A})$ не равна в точности нулю, но может быть сделана сколь угодно малой. Соответственно сколь угодно малой может быть сделана вероятность ошибочного декодирования. Как будет видно из доказательства, чем меньше допустимая вероятность ошибочного декодирования, тем сложнее должен быть код.

Идея доказательства основной теоремы кодирования Шеннона для дискретного канала основана на случайном кодировании. Рассмотрим некоторое множество кодов и определим среднюю (по всему этому множеству кодов) вероятность ошибочного декодирования, которая в условиях теоремы окажется меньше наперед заданной величины ε. Отсюда следует, что в рассмотренном множестве кодов содержатся такие, для которых вероятность ошибочного декодирования также меньше ε, чем и будет завершено доказательство.

Пусть при некотором ансамбле B_0 входных сигналов дискретного канала обеспечивается пропускная способность канала:

$$C = I'(B_0, \hat{B}_0) = H'(B_0) - H'(B_0|\hat{B}_0). \tag{3.75}$$

В соответствии с теоремой об асимптотической равновероятности (см. § 2.8) число типичных последовательностей входных сигналов достаточно большой длительности T (содержащих достаточно большое число символов n) равно $N_{\text{тип}}(B_0)=2^{TH'(B_0)}$.

Пусть передаче подлежат сообщения, выдаваемые дискретным источником, производительность которого меньше пропускной способности канала: $H'(A)<C=H'(B_0)-H'(B_0|\hat{B}_0)$. Поскольку ненадежность $H'(B_0|\hat{B}_0)\geqslant 0$, то

$$H'(A) < H'(B_0). \tag{3.76}$$

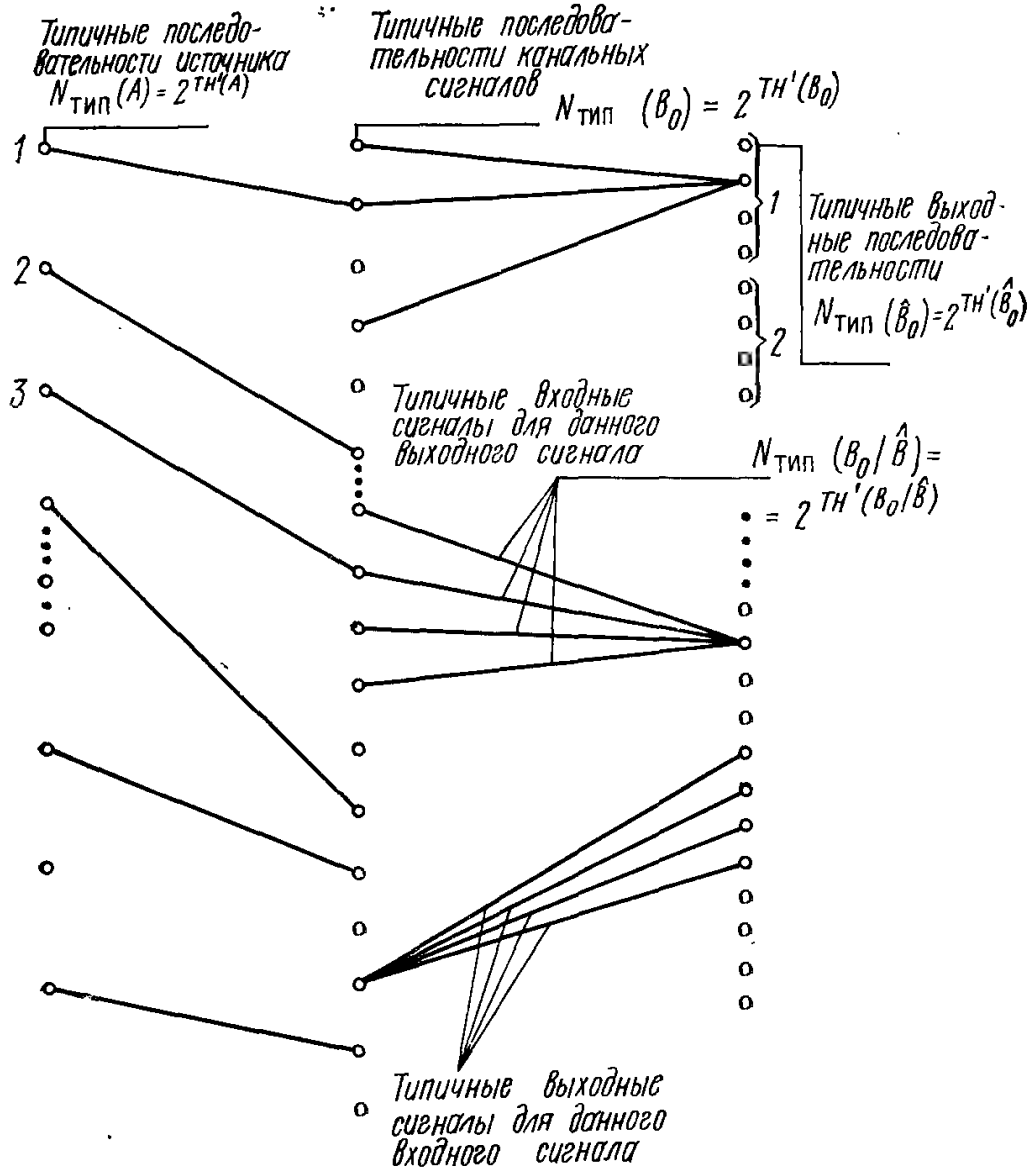


Рис. 3.12. Иллюстрация идеи случайного кодирования типичных последовательностей источника и возможных переходов типичных входных последовательностей в канале с шумами

Число типичных последовательностей источника достаточно большой длительности T равно $N_{тип}(A) = 2^{TH'(A)}$. Будем каждой типичной последовательности источника ставить в соответствие одну из типичных последовательностей канальных сигналов B_0 (рис. 3.12). Нетипичные же последовательности сообщений длительности T (если источник все же выдаст какую-нибудь из них) передавать вовсе не будем, соглашаясь с тем, что каждая такая последовательность будет принята ошибочно. Вероятность появления такой последовательности меньше наперед заданной величины δ_1.

Выполним указанное кодирование (сопоставление последовательностей источника и канала) всеми возможными способами. Очевидно, что число возможных кодов M равно числу размещений из $N_{тип}(B_0)$ элементов по $N_{тип}(A)$. Каждому из них соответствует своя вероятность ошибки p_k. Средняя по всем этим способам вероятность ошибки

$$\bar{p} = \frac{1}{M}\sum_{k=1}^{M} p_k.$$

Средияя вероятность ошибочного приема типичной последовательности источника может быть получена в предположении случайного выбора типичной кодовой последовательиости сообщений той же длительности.

При достаточно большом T вероятность того, что какая-то типичная последовательность канальных сигналов использована при кодировании, в силу равиовероятиости выбора, с учетом (3.76)

$$p_{\text{исп}} = N_{\text{тип}}(A)/N_{\text{тип}}(B_0) = 2^{-T[H'(B_0)-H'(A)]} \ll 1. \tag{3.77}$$

Пусть принят некоторый выходной сигнал (последовательность символов) $\hat{b}$ длительностью T. Он мог, в принципе, получиться при передаче любого входиого сигнала b. Среди них существует $N_{\text{тип}}(B_0|\hat{B}) = 2^{T[H'(B_0|B)]}$ типичных, а вероятность того, что в действительности передавался некоторый сигнал, не входящий в число этих «условно типичных», при достаточно большом T меньше любого наперед заданного положительного числа δ_2.

Установим следующее простое правило декодирования. При приходе некоторой реализации сигнала $\hat{b}$ будем считать, что передавалось сообщение a, которое при кодировании было сопоставлено с одним из $N_{\text{тип}}(B_0|\hat{B})$ условно типичных входных сигналов. Если таких сообщений несколько, то выбираем любое из них наугад.

Ошибочное декодирование возможно в трех случаях: если среди $N_{\text{тип}}(B_0|\hat{B})$ последовательностей сигналов имеются две или более использованные при кодировании, если передавалось нетипичное сообщение и если произошел нетипичный переход переданного сигнала b в принятый $\hat{b}$. Последиие два события имеют вероятности δ_1 и δ_2.

Таким образом, среднюю по всем кодам вероятность ошибочного декодирования $\bar{p}$ сигнала длительностью T можно оценить неравенством Буля[1]

$$\bar{p} \leqslant \bar{p}_1 + \delta_1 + \delta_2, \tag{3.78}$$

где $\bar{p}_1$ — средняи по всем кодам вероятность того, что среди $N_{\text{тип}}(B_0|\hat{B})$ входных символов при кодировании использовано более одиого.

На основании (3.77) и (3.75)

$$\bar{p}_1 = 1 - (1 - p_{\text{исп}})^{N_{\text{тип}}(B_0|\hat{B})-1} < 1 - (1 - p_{\text{исп}})^{N_{\text{тип}}(B_0|\hat{B})} \leqslant$$

$$\leqslant 1 - 1 + p_{\text{исп}} N_{\text{тип}}(B_0|\hat{B}) = 2^{-T[H'(B_0)-H'(A)]} 2^{TH'(B_0|\hat{B})} =$$

$$= 2^{-T[H'(B_0)-H'(B_0|\hat{B})-H'(A)]} = 2^{-T[C-H'(A)]}, \tag{3.79}$$

и так как по условию теоремы $C - H'(A) > 0$, то с увеличением T вероятность $\bar{p}_1$ стремится к 0. Так как при этом δ_1 и δ_2 также стремятся к нулю, то $\bar{p}$ при достаточно большом T будет меньше наперед заданной положительной величины ε. Поскольку среди M различных кодов существует хотя бы один, при котором вероятность ошибочного декодирования не превышает среднего значения, то первая часть теоремы доказана.

Вторая же часть непосредственно следует из того установленного выше факта, что условие $H'(A) < C$ является необходимым для точной передачи сообщения.

В частном случае канала без помех, когда $C = \log m$, доказанная теорема сводится к утверждению, что *сообщения источника можно закодировать m-ичным кодом, если на каждое элементарное сообщение затрачивать в среднем* $\bar{n} = v/v_{\text{н}} > H(A)/\log m$ *символов* или *для двоичного кода* $\bar{n} > H(A)$. Это совпадает с теоремой, доказанной в § 2.9.

---

[1] Согласно этому неравенству вероятность P_n того, что произойдет хотя бы одно из n событий $A_1, A_2, \ldots, A_n$, меньше или равна сумме вероятностей этих событий: $P_n \leqslant \sum_{k=1}^{n} P(A_k)$. Это иеравенство переходит в равенство, если события A_k попарно несовместимы.

Теорема кодирования Шеннона справедлива и при передаче дискретных сообщений по непрерывному каналу. В этом случае под кодированием понимают отбор некоторого количества реализаций $u(t)$ входного сигнала на интервале T и сопоставление с каждой из них последовательности элементарных сообщений, выдаваемой источником за тот же интервал T.

Доказательство этой теоремы для непрерывного канала аналогично доказательству для дискретного канала. Оно довольно сложно и поэтому здесь не приводится. Заметим лишь, что вероятность ошибочного декодирования сигнала длительностью T оценивается формулой, аналогичной (3.79):

$$p_{\text{ош}}(T) \leqslant \exp\{-\eta\,[C-H'(A)]\,T\}, \qquad (3.80)$$

где η — коэффициент, определяющий скорость экспоненциального убывания вероятности ошибки с увеличением T. В общем случае η зависит от свойств канала, а также от T, но при $H'(A)<C$ для всех каналов $\eta>0$, так что с увеличением T принципиально можно получить сколь угодно малую вероятность ошибочного приема[1].

Подчеркнем важный результат, непосредственно следующий из формул (3.79) и (3.80): верность связи тем выше, чем больше T, т. е. чем длиннее кодируемый отрезок сообщения (а следовательно, и больше задержка при приеме информации), и чем менее эффективно используется пропускная способность канала (чем больше разность $C-H'(A)$, определяющая «запас пропускной способности» канала). Итак, существует возможность обмена между верностью, задержкой и эффективностью системы. С увеличением T существенно возрастает сложность кодирования и декодирования (число операций, число элементов и стоимость аппаратуры). Поэтому практически чаще всего предпочитают иметь умеренное значение задержек T, которые, кстати, не во всех системах связи можно произвольно увеличивать, и добиваются повышения верности за счет менее полного использования пропускной способности канала.

Примеры практического применения кодирования для повышения верности передачи информации будут приведены в последующих главах.

ГЛАВА ЧЕТВЕРТАЯ

ПЕРЕДАЧА ДИСКРЕТНЫХ СООБЩЕНИЙ В НЕПРЕРЫВНЫХ КАНАЛАХ

4.1. ПРИЕМ СИГНАЛОВ КАК СТАТИСТИЧЕСКАЯ ЗАДАЧА

В этой главе основное внимание уделено модему, в частности демодулятору. На его вход (см. рис. 1.5, с. 20) поступает сигнал

[1] Следует подчеркнуть, что здесь идет речь о передаче дискретных сообщений в непрерывном канале. Для того чтобы сформулировать аналогичную теорему для источника непрерывных сообщений, необходимо, прежде всего, определить его производительность. Это будет сделано в гл. 6.

с выхода непрерывного канала, искаженный аддитивными и мультипликативными помехами[1]. На выходе же демодулятора возникает дискретный сигнал, т. е. последовательность кодовых символов. Обычно некоторый отрезок (элемент) непрерывного сигнала преобразуется модемом в один кодовый символ (поэлементный прием). Если бы этот кодовый символ всегда совпадал с передаваемым (поступившим на вход модулятора), то связь была бы безошибочной. Но как уже известно, помехи приводят к невозможности с абсолютной достоверностью восстановить по принятому сигналу переданный кодовый символ.

Каждый демодулятор математически описывается законом, по которому поступивший на его вход непрерывный сигнал превращается в кодовый символ. Этот закон называется *правилом решения*, или *решающей схемой*. Демодуляторы с различными правилами решения будут выдавать, вообще говоря, различные решения, из которых одни будут верными, а другие ошибочными.

Будем полагать, что свойства источника сообщения и кодера известны. Кроме того, известен модулятор, т. е. задано, какая реализация элемента сигнала соответствует тому или иному кодовому символу, а также задана математическая модель непрерывного канала. Требуется определить, каков должен быть демодулятор (правило решения), чтобы обеспечить оптимальное (т. е. наилучшее из возможных) качество приема.

Такая задача была впервые поставлена и решена (для гауссовского канала) в 1946 г. выдающимся советским ученым В. А. Котельниковым. В этой постановке качество оценивалось вероятностью правильного приема символа. Максимум этой вероятности при заданном виде модуляции В. А. Котельников назвал *потенциальной помехоустойчивостью,* а демодулятор, обеспечивающий этот максимум, — *идеальным приемником.* Из этого определения следует, что ни в одном реальном демодуляторе вероятность правильного приема символа не может быть больше, чем в идеальном приемнике.

На первый взгляд принцип оценки качества приема вероятностью правильного приема символа кажется вполне естественным и даже единственно возможным. Ниже будет показано, что это не всегда так и что существуют и другие критерии качества, применимые в тех или иных частных случаях.

Ознакомимся несколько подробнее со статистическим подходом к задаче приема дискретных сообщений на фоне шумов. Пусть при передаче дискретных сообщений, закодированных кодом с основанием m, используются реализации сигнала $u_i(t)$ $(0 \leqslant t \leqslant T)$, соответствующие кодовым символам b_i $(i=1, 2, 3, \ldots, m)$. В течение

---

[1] Обычно в приемных устройствах демодулятору предшествуют усилители и преобразователи частоты. В данной главе все они считаются включенными в состав канала. Между прочим, именно они являются в ряде случаев основными источниками аддитивных помех канала.

тактового интервала[1] $0<t\leqslant T$ на вход приемного устройства поступает колебание $z(t)$, которое вследствие искажений и помех в канале не совпадает в точности ни с одним из сигналов $u_i(t)$. Следовательно, в этом случае приемное устройство должно выбрать одну из m возможных взаимоисключающих (альтернативных) гипотез:

$$\left.\begin{array}{l}\text{передавался кодовый символ } b_1, \text{ т. е. сигнал } u_1(t);\\ \text{передавался кодовый символ } b_2, \text{ т. е. сигнал } u_2(t);\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \text{передавался кодовый символ } b_m, \text{ т. е. сигнал } u_m(t).\end{array}\right\} \quad (4.1)$$

Совокупность всех возможных реализаций $z(t)$ можно интерпретировать точками в пространстве Z принимаемых финитных сигналов. Обычно оно является бесконечномерным пространством Гильберта или, с некоторыми (приемлемыми для практики) оговорками, многомерным пространством Евклида. Простоты ради будем графически изображать реализации принимаемых сигналов $s_i(t)$ и помехи $n(t)$ длительностью T точками на плоскости (рис. 4.1) или соответствующими векторами, откладываемыми от начала координат 0. Если правило решения выбрано, то это означает, что каждой точке пространства принимаемых колебаний (концу вектора) $\mathbf{z}=\mathbf{s}+\mathbf{n}$ приписывается одна из m гипотез, т. е. определенный передаваемый кодовый символ b_i. Пространство принимаемых сигналов окажется при этом разбитым на m непересекающихся областей $\hat{B}_i$, каждая из которых соответствует принятию определенной гипотезы. В такой трактовке различные приемные устройства отличаются друг от друга способом разбиения пространства сигналов на области $\hat{B}_i$, т. е. правилом принятия решения[2]. Возможное разбиение схематически показано на рис. 4.1.

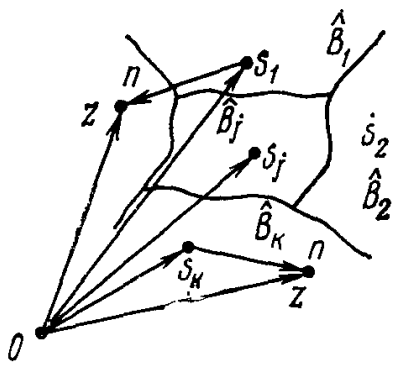


Рис. 4.1. Разбиения пространства принимаемых колебаний на непересекающиеся области

Пусть на интервале 0—T принимается колебание

$$z(t)=s_j(t)+n(t), \quad (4.2)$$

где $s_j(t)$ — полезный сигнал в месте приема, прошедший канал связи, а $n(t)$ — реализация аддитивной помехи.

---

[1] Начало этого интервала для удобства совместим с началом координат. В принципе, интервал анализа не всегда совпадает с тактовым интервалом T. Отрезок сигнала на тактовом интервале называют элементом сигиала.

[2] В математической теории связи это разбиение и называют решающей схемой. Заметим, что в некоторых случаях пользуются решающей схемой со *стиранием*, или *отказом от решения*. Это значит, что m областей не охватывают всего пространства сигналов z, и если приходящий сигнал не попадает ни в одну из областей, то никакого решения ие принимается.

Если помех нет, возможные значения $z(t)$ изображаются точками $\mathbf{s}_j (j=1, 2, 3, \ldots, m)$. При наличии помехи и при передаче сигнала с номером j точка принимаемого колебания $\mathbf{z}$ отклоняется от точки $\mathbf{s}_j$. Обычно область $\hat{B}_j$ содержит точку $\mathbf{s}_j$. В тех случаях, когда помеха не выводит точку $\mathbf{z}$ за пределы области $\hat{B}_j$, решение оказывается верным. В противном случае возникает ошибка. Очевидно, изменяя границы между областями, можно влиять на вероятность ошибочного приема отдельных передаваемых символов. Например, если в разбиении, показанном на рис. 4.1, расширить область $\hat{B}_j$, изменив ее границы с областью $\hat{B}_k$, то уменьшится вероятность ошибочного приема символа $\hat{b}_k$ вместо передаваемого символа b_j. Однако в этом случае возрастает вероятность ошибочного приема $\hat{b}_j$ при передаваемом b_k. Очевидно, всегда существует такое расположение областей, которое в определенном смысле лучше всякого другого.

Если задан критерий качества, то наилучшее разбиение пространства принимаемых сигналов (оптимальная решающая схема приемного устройства) находится методом теории статистических решений.

4.2. КРИТЕРИИ КАЧЕСТВА ПРИЕМА ДИСКРЕТНЫХ СООБЩЕНИЙ

Рассмотрим сначала уже упоминавшийся и широко распространенный критерий Котельникова, или идеального наблюдателя, согласно которому качество демодулятора оценивается безусловной вероятностью правильного приема символа.

Будем вначале полагать, что пространство передаваемых и принимаемых сигналов является конечномерным евклидовым. Это может быть, например, пространство финитных сигналов, представляемых конечной тригонометрической суммой. В дальнейшем это ограничение будет отброшено.

В n-мерном пространстве случайный сигнал $z(t)$ характеризуется n-мерной плотностью вероятностей вектора $\mathbf{z}$: $w(\mathbf{z})$. Ее можно рассматривать как плотность вероятности коэффициентов разложения $z(t)$ по любому ортонормированному базису. Если передается некоторый символ b_i, т. е. посылается сигнал $u_i(t)$, то можно определить условную n-мерную плотность вероятности $w\{\mathbf{z}|b_i\}$.

Пусть на вход демодулятора в течение тактового интервала $0—T$ приходит некоторый элемент сигнала $z(t)$. Предположим, что демодулятор принимает при этом решение, что передан символ b_i. Вероятность того, что это решение правильно, очевидно, равна условной вероятности того, что действительно передавался символ b_i при условии прихода реализации элемента сигнала $z(t)$, $P(b_i|z)$. Ее называют обычно *апостериорной* вероятностью символа b_i (т. е. вероятностью, определенной после опыта, заключающегося в наблюдении и анализе сигнала $z(t)$).

Очевидно, что вероятность правильного приема будет максимальной в такой решающей схеме, которая относит всякую реализацию элемента приходящего сигнала $z(t)$ к той области $\hat{B}_i$, для

которой апостериорная вероятность $P(b_i|z)$ максимальна. Другими словами, критерий идеального наблюдателя обеспечивается решающей схемой, построенной по правилу *максимума апостериорной вероятности* — решение b_i принимается в том случае, если выполняется система из $m-1$ неравенств:

$$P(b_i \mid \mathbf{z}) > P(b_j \mid \mathbf{z}), \quad j = 1, \ldots, m, \quad j \neq i. \tag{4.3}$$

Для сокращения запишем это правило в такой форме:

$$\max_i P(b_i \mid \mathbf{z}). \tag{4.4}$$

Согласно известной формуле Байеса

$$P(b_j \mid \mathbf{z}) = [P(b_j) w(\mathbf{z} \mid b_j)/w(\mathbf{z})], \tag{4.5}$$

где $P(b_j)$ — *априорная* вероятность передачи символа b_j (т. е. та вероятность, которая имеет место до наблюдения и анализа $z(t)$, определяемая статистикой источника сообщения и правилом кодирования).

Подставив (4.4) в (4.2) и (4.3) и учитывая, что $w(\mathbf{z})$ — безусловная плотность вероятности, не являющаяся функцией j, можно записать правило решения по критерию идеального наблюдателя в следующей форме:

$$P(b_i) w(\mathbf{z} \mid b_i) > P(b_j) w(\mathbf{z} \mid b_j), \quad j = 1, \ldots, m; \; j \neq i \tag{4.6}$$

или сокращенно

$$\max_i [P(b_i) w(\mathbf{z} \mid b_i)]. \tag{4.7}$$

Для построения такой решающей схемы необходимо знать априорные вероятности символов $P(b_j)$, а также свойства модулятора и канала, определяющие условные плотности $w(\mathbf{z}|b_j)$.

Правило (4.6) можно записать иначе — решение о том, что передавался символ b_i, должно приниматься, если для всех $j \neq i$ выполняется $m-1$ неравенств:

$$\frac{w(\mathbf{z} \mid b_i)}{w(\mathbf{z} \mid b_j)} > \frac{P(b_j)}{P(b_i)}. \tag{4.8}$$

Отношение в левой части этого неравенства называется *отношением правдоподобия* двух гипотез о том, что передавался символ b_i, и о том, что передавался символ b_j. Его обозначают $\Lambda_{i,j}$.

В случае, когда все m символов передаются равновероятно, т. е. $P(b_i) = P(b_j) = 1/m$, правило (4.8) упрощается:

$$\Lambda_{i,j} > 1. \tag{4.9}$$

Иногда вводят в рассмотрение помимо m гипотез о передаче символов b_i $(i = 1, 2, \ldots, m)$ еще одну «нулевую» гипотезу о том, что никакой сигнал не передавался, т. е. $z(t)$ — чистая помеха.

Отношение правдоподобия $\frac{w(\mathbf{z} \mid b_i)}{w(z \mid 0)} = \Lambda_{i,\,0}$ обычно обозначают просто Λ_i, тогда правило (4.9) можно записать так:

$$\Lambda_i > \Lambda_j \text{ при всех } j \neq i, \tag{4.10}$$

или, короче, $$\max_i [\Lambda_i]. \tag{4.11}$$

Такое *правило максимума правдоподобия* реализует критерий идеального наблюдателя только при том условии, что все символы передаются равновероятно[1].

Как уже отмечалось, критерий идеального наблюдателя не является единственным разумным критерием оптимальности решающей схемы. Дело в том, что во многих случаях различные ошибки приводят к различным последствиям. Например, в системе автоматической пожарной сигнализации опаснее не обнаружить сигнал о пожаре, нежели получить ложную тревогу, когда в действительности пожара нет.

Учет последствий ошибок различного рода (связанных с передачей различных символов) приводит к обобщению критерия идеального наблюдателя, известного под названием критерия *минимального среднего риска* (или байесовского критерия). Введем некоторые понятия. Если при передаче символа b_i принят символ $\hat{b}_j$, то при $j \neq i$ имеет место ошибка. Чтобы учесть неравноценность различных ошибок, будем с каждой парой символов $\hat{b}_j$ и b_i связывать некоторую численную величину, называемую «потерей», обозначив ее L_{ij}. Величина «потери» зависит, таким образом, от того, какой символ $\hat{b}_j$ принят вместо переданного b_i. Правильному приему при этом обычно приписывается нулевая «потеря». Значения L_{ij} определяются в каждом конкретном случае важностью правильного приема данного элемента сигнала и величиной опасности различных ошибок.

Так как при передаче символа b_i символы $\hat{b}_j$ появляются с определенными вероятностями как реализации некоторой дискретной случайной величины, можно говорить об условном математическом ожидании величины «потери» при передаче конкретного символа b_i. Назовем это условное математическое ожидание *условным риском:*

$$R_i = \sum_{j=1}^{m} P(\hat{b}_j \mid b_i) L_{ij} = \sum_{j=1}^{m} L_{ij} \int_{\hat{B}_j} w(\mathbf{z} \mid b_i)\, d\mathbf{z}. \tag{4.12}$$

Интеграл в (4.12) берется по области $\hat{B}_j$ решающей схемы и представляет вероятность того, что сигнал $z(t)$ попал в эту область, если передавался символ b_i. Усреднив условный риск R_i по всем символам b_i, получим величину, называемую *средним риском:*

$$R_{\text{ср}} = \sum_{i=1}^{m} \sum_{j=1}^{m} P(b_i) L_{ij} \int_{\hat{B}_j} w(\mathbf{z} \mid b_i)\, d\mathbf{z}. \tag{4.13}$$

---

[1] Вместо неравенств (4.10) можно было бы просто записать $w(\mathbf{z}|b_i) > w(\mathbf{z}|b_j)$. Сравнение отношений правдоподобия вместо сравнения условных плотностей вероятностей вызвано тем, что понятие отношения правдоподобия можно в известном смысле распространить и на сигналы из бесконечномерного гильбертовского пространства, для которых понятие плотностей вероятности $w(\mathbf{z})$, $w(\mathbf{z}|b_i)$ теряет смысл. Как это делается, будет показано в § 4.3.

Критерий *минимального среднего риска* заключается в том, что оптимальной считается решающая схема, обеспечивающая наименьшее значение среднего риска $R_{ср}$. Приемник, работающий по такому критерию, называют байесовским.

Из (4.13) видно, что при использовании этого критерия нужно помимо априорных вероятностей $P(b_j)$ передачи отдельных символов знать и величины потерь L_{ij}. Заметим, что если считать все ошибки равноценными ($L_{ij}=\text{const}$ при $j\neq i$ и $L_{ii}=0$), то критерий минимального среднего риска совпадает с критерием идеального наблюдателя, а байесовский приемник совпадает с идеальным приемником Котельникова. В общем же случае в оптимальном байесовском приемнике чаще будут возникать ошибки, связанные с малыми потерями, и реже — с большими потерями.

Как видно из (4.13), недостаток, связаиный с необходимостью знать априорные вероятности передаваемых символов, присущ и критерию минимального среднего риска. Кроме того, его применение усложняется трудностью объективного установления значений потерь. Тем не менее этот критерий как обобщение критерия идеального наблюдателя целесообразно использовать, если последствия различных ошибок неравноценны.

Ситуация, в которой практически невозможно определить априорную вероятность передачи отдельных элементарных сообщений, а последствия ошибок разного рода неодинаковы, особенно типична для радиолокации, когда приемник, анализируя принимаемое колебание $z(t)$ (отраженный сигнал плюс помеха), должен определить, имеется в данном направлении и на данном расстоянии объект наблюдения (цель) или нет. Как правило, априорная вероятность $p(1)$ наличия отраженного от цели сигнала (передачи символа «1») заранее неизвестна. Последствия двух родов ошибок — ложной тревоги (приемник фиксирует, что цель существует, в то время как в действительности ее нет) и пропуска цели (приемник отмечает отсутствие цели, в то время как фактически она имеется) — неравноценны.

В этой и других сходных ситуациях чаще всего пользуются критерием приема, известным под названием *критерия Неймана — Пирсона.* Суть его заключается в том, что решающая схема считается оптимальной, если при заданной вероятности ложной тревоги $p_{л.т}$ обеспечивается минимальная вероятность пропуска цели $p_{пр}$.

Очевидно, что можно различными способами разбить пространство принимаемых колебаний $z(t)$ на две области: $\hat{B}_0$ (область решения об отсутствии цели) и $\hat{B}_1$ (о наличии цели) — так, чтобы вероятность ложной тревоги

$$P_{л.т}=\int\limits_{\hat{B}_1} w(\mathbf{z}\mid 0)\,d\mathbf{z} \tag{4.14}$$

равнялась заданной величине. Поскольку в локации символ «0» (отсутствие цели) передается паузой, то $w(\mathbf{z}\mid 0)$ — это плотность распределения помехи. Следовательно, вероятность ложной тревоги определяется вероятностными характеристиками помехи и выбором области $\hat{B}_1$. Но от выбора области $\hat{B}_1$ зависит и вероятность пропуска цели:

$$p_{пр}=\int\limits_{\hat{B}_0} w(\mathbf{z}\mid 1)\,d\mathbf{z}=1-\int\limits_{\hat{B}_1} w(\mathbf{z}\mid 1)\,d\mathbf{z}. \tag{4.15}$$

Интегралы в (4.14), (4.15) и в аналогичных других формулах, взятые по векторной переменной, очевидно, n-кратные.

Минимизация (4.15) при заданной величине (4.14) достигается, если решение о наличии цели принимается при выполнении неравенства

$$\Lambda_1=w(\mathbf{z}\mid 1)\,/\,w(\mathbf{z}\mid 0)>\lambda, \tag{4.16}$$

где λ — пороговый уровень, определяемый заданной вероятностью ложной тревоги $p_{л.т}$.

Существуют и другие критерии качества приема, не требующие знания априорных вероятностей символов.

В технике связи (как и в ряде других систем передачи информации) преимущественное применение находит правило максимального правдоподобия (4.11). В том случае, когда все символы передаются равновероятно, правило максимального правдоподобия реализует критерий идеального наблюдателя. Однако очень часто это правило решения применяют и при неизвестных или известных, но не одинаковых, априорных вероятностях символов. Конечно, оно не обеспечивает в этих случаях максимума вероятности правильного приема. Изменив решающую схему на схему, построенную по правилу максимальной апостериорной вероятности (4.4), реализующему критерий идеального наблюдателя, можно было бы уменьшить вероятность ошибок. При этом, очевидно, пришлось бы сократить области приема маловероятных символов и расширить области высоковероятных символов. В результате редко передаваемые символы принимались бы менее надежно, нежели часто передаваемые символы. Но редкие символы несут больше информации, чем частые [см. ф-лу (2.119)]. Поэтому переход от правила максимального правдоподобия к правилу максимальной апостериорной вероятности, хотя и уменьшает безусловную вероятность ошибки, может привести к увеличению потери информации при демодуляции[1]. Легко показать, что правило максимального правдоподобия реализует критерий минимума среднего риска (4.13), если положить $L_{ij}=0$ при $i=j$ и $L_{ij}=1/P(b_i)$ при $i\neq j$.

Заметим также, что для большей части дискретных систем связи различие между правилами максимальной апостериорной вероятности и максимального правдоподобия невелико. Это объясняется тем, что при достаточно эффективном кодировании вероятности передачи символов почти одинаковы.

Вследствие сказанного будем в дальнейшем пользоваться правилом максимального правдоподобия и решающую схему, реализующую правило (4.11), будем условно называть оптимальной.

4.3. ОПТИМАЛЬНЫЕ АЛГОРИТМЫ ПРИЕМА ПРИ ПОЛНОСТЬЮ ИЗВЕСТНЫХ СИГНАЛАХ (КОГЕРЕНТНЫЙ ПРИЕМ)

Предположим, что все искажения в канале строго детерминированы и случайным является только гауссовский аддитивный шум $N(t)$, который будем вначале полагать белым, со спектральной плотностью N_0. Это значит, что при передаче сигнала $u_i(t)$ (символа b_i, $i=1, 2, \ldots, m$) приходящий сигнал можно описать моделью (3.28):

$$z(t)=s_i(t)+n(t), \quad (0\leqslant t\leqslant T), \tag{4.17}$$

где все $s_i(t)=ku_i(t-\tau)$ $(i=1, 2, \ldots, m)$ известны. Неизвестны лишь реализация помехи и индекс i действительно переданного сигнала, который и должна определить решающая схема.

[1] Как показал Н. И. Клюев, правило (4.11) оптимально по критерию максимума частного количества информации в принятом сигнале относительно переданного символа.

Будем также считать, что все $s_i(t)$ являются финитными сигналами, длительность которых T. Это имеет место, если передаваемые сигналы $u_i(t)$ финитны и имеют одинаковую длительность (система синхронная), а в канале нет ни многолучевого распространения, ни линейных искажений, вызывающих растяжение сигнала (либо они скорректированы).

В дальнейшем будем везде полагать, что в системе обеспечена надежная тактовая синхронизация, т. е. границы тактового интервала, на котором приходит сигнал $s(t)$, известны точно. Вопросы синхронизации весьма существенны при реализации оптимальных демодуляторов и синхронных систем связи вообще, но они выходят за пределы данного курса. Момент начала посылки $s(t)$ примем за нуль.

Определим в этих условиях алгоритм работы оптимального (т. е. основанного на правиле максимального правдоподобия) демодулятора, анализирующего сигнал на тактовом интервале $0—T$. С этой целью необходимо найти отношения правдоподобия для всех m возможных сигналов относительно нулевой гипотезы ($s(t)=0$; $z(t)=n(t)$).

Задача затрудняется тем, что ширина спектра сигнала бесконечна (поскольку он финитный), а поэтому пространство сигналов бесконечномерное $L_2(T)$. Для таких сигналов (или бесконечномерных векторов), как уже отмечалось, не существует плотности вероятностей. Однако существуют n-мерные плотности вероятностей для любых n сечений сигнала (см. § 2.1).

Заменим вначале белый шум квазибелым, имеющим ту же одностороннюю спектральную плотность мощности N_0, но только в некоторой полосе частот $F=n/2T$, где $n\gg 1$. Рассмотрим вначале нулевую гипотезу, т. е. будем считать, что $Z(t)$ — шум. Возьмем на тактовом интервале n равноотстоящих сечений через $\Delta t=1/2F=T/n$. Отсчеты $Z_1,\ldots,Z_n$ в этих сечениях для квазибелого гауссовского шума независимы в соответствии с (2.49). Поэтому n-мерная плотность вероятностей для взятых отсчетов

$$w(z_1, z_2, \ldots, z_n; t_1, t_2, \ldots, t_n \mid 0) =$$

$$= \frac{1}{(\sqrt{2\pi}\,\sigma)^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{n} z^2(t_k)\right], \tag{4.18}$$

где $\sigma^2=N_0F$ — дисперсия (мощность) квазибелого шума.

При гипотезе, что передавался символ b_i, $n(t)=z(t)-s_i(t)$. Следовательно, условная n-мерная плотность вероятности сечений $Z(t)$ определится такой же формулой, как и (4.18), если $z(t_k)$ заменить разностью $z(t_k)-s_i(t_k)$:

$$w(z_1, \ldots, z_n; t_1, \ldots, t_n \mid b_i) =$$

$$= \frac{1}{(\sqrt{2\pi}\,\sigma)^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n} [z(t_k)-s_i(t_k)]^2\right\}. \tag{4.19}$$

Отношение правдоподобия для сигнала s_i (относительно нулевой гипотезы), вычисленное для n сечений:

$$\Lambda_i^{[n]} = \frac{w_n(z_1, z_2, \ldots, z_n; t_1, t_2, \ldots, t_n \mid b_i)}{w_n(z_1, z_2, \ldots, z_n; t_1, t_2, \ldots, t_n \mid 0)} =$$

$$= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{k=1}^{n} [z(t_k) - s_i(t_k)]^2 + \frac{1}{2\sigma^2} \sum z^2(t_k) \right\}. \quad (4.20)$$

Заменим дисперсию σ^2 ее выражением:

$\sigma^2 = N_0 F = N_0/(2\Delta t)$.

$$\text{Тогда } \Lambda_i^{[n]} = \exp\left\{ -\frac{1}{N_0} \sum_{k=1}^{n} [z(t_k) - s_i(t_k)]^2 \Delta t + \frac{1}{N_0} \sum_{k=1}^{n} z^2(t_k) \Delta t \right\}. \quad (4.21)$$

По правилу максимума правдоподобия в случае квазибелого шума решающая схема должна выбирать значение i, обеспечивающее максимум $\Lambda_i^{[n]}$. Вместо максимума Λ_i можно отыскивать максимум его логарифма:

$$\ln \Lambda_i^{[n]} = -\frac{1}{N_0} \sum_{k=1}^{n} [z(t_k) - s_i(t_k)]^2 \Delta t + \frac{1}{N_0} \sum_{k=1}^{n} z^2(t_k) \Delta t. \quad (4.22)$$

Заметим, что второй член в (4.22) не зависит от i и его можно при сравнении не учитывать. Тогда правило решения о том, что передавался сигнал $s_i(t)$, можно сформулировать следующим образом:

$$\sum_{k=1}^{n} [z(t_k) - s_i(t_k)]^2 \leqslant \sum_{k=1}^{n} [z(t_k) - s_j(t_k)]^2, \quad (j = 0, \ldots, m-1). \quad (4.23)$$

В n-мерном евклидовом пространстве $\sqrt{\sum_{k=1}^{n} [z(t) - s_i(t)]^2}$ определяет норму разности векторов $\mathbf{z}$ и $\mathbf{s}_i$, или расстояние между ними. Поэтому алгоритм (4.23) можно записать в виде

$$\| \mathbf{z} - \mathbf{s}_i \| < \| \mathbf{z} - \mathbf{s}_j \|, \quad (j = 0, \ldots, m-1)$$

и придать ему простую геометрическую интерпретацию: оптимальный демодулятор должен регистрировать тот из сигналов $s_i(t)$ (соответствующий символу b_i), который «ближе» к принятому колебанию $z(t)$. В качестве примера на рис. 4.2 показано оптимальное разбиение двумерного пространства принимаемых сигналов $z(t)$ при передаче двоичных сигналов $s_0(t)$ и $s_1(t)$. Области принятия решения в пользу символов b_0 и b_1 расположены по обе стороны от

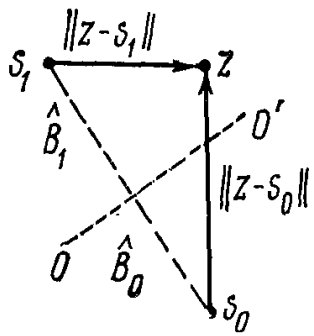


Рис. 4.2. Оптимальное разбиение пространства принимаемых колебаний при двоичном коде и точно известных сигналах

прямой 0—0′, перпендикулярной отрезку, соединяющему точки сигналов, и делящей его пополам.

Преобразуем (4.22), раскрыв скобки и произведя сокращения:

$$\ln \Lambda_i^{[n]} = \frac{2}{N_0} \sum_{k=1}^{n} z(t_k)\, s_i(t_k)\, \Delta t - \frac{1}{N_0} \sum_{k=1}^{n} s_i^2(t_k)\, \Delta t. \qquad (4.24)$$

Вернемся теперь к исходной задаче для белого шума. С этой целью будем расширять полосу F. Тогда число сечений n будет стремиться к бесконечности, а Δt — к нулю. Суммы в (4.24) обратятся в интегралы, и логарифм отношения правдоподобия определится как

$$\ln \Lambda_i = \lim_{n \to \infty} \ln \Lambda_i^{[n]} = \frac{2}{N_0} \int_0^T z(t)\, s_i(t)\, dt - \frac{1}{N_0} \int_0^T s_i^2(t)\, dt, \qquad (4.25)$$

а алгоритм решения о передаче $s_i(t)$ примет вид

$$\int_0^T z(t)\, s_i(t)\, dt - 0{,}5\, E_i \geqslant \int_0^T z(t)\, s_j(t)\, dt - 0{,}5\, E_j, \quad j = 0, \cdots, m-1, \qquad (4.26)$$

где E_j — энергия ожидаемого сигнала $s_j(t)$:

$$E_j = \int_0^T s_j^2(t)\, dt. \qquad (4.27)$$

Устройство, непосредственно вычисляющее скалярное произведение,

$$(\mathbf{z},\ \mathbf{s}_i) = \int_0^T z(t)\, s_i(t)\, dt \qquad (4.28)$$

называют активным фильтром, или коррелятором, поэтому приемник, реализующий алгоритм (4.26), называют *корреляционным.*

На рис. 4.3 показана структурная схема приемного устройства, работающего в соответствии с (4.26). Здесь блоки $\times$ — перемножители; Γ_i — генераторы опорных сигналов $s_i(t)$; $\int$ — интеграторы, «—» — вычитающие устройства; РУ — решающее устройство, определяющее в моменты времени, кратные T (при замыкании ключа), номер ветви с максимальным сигналом.

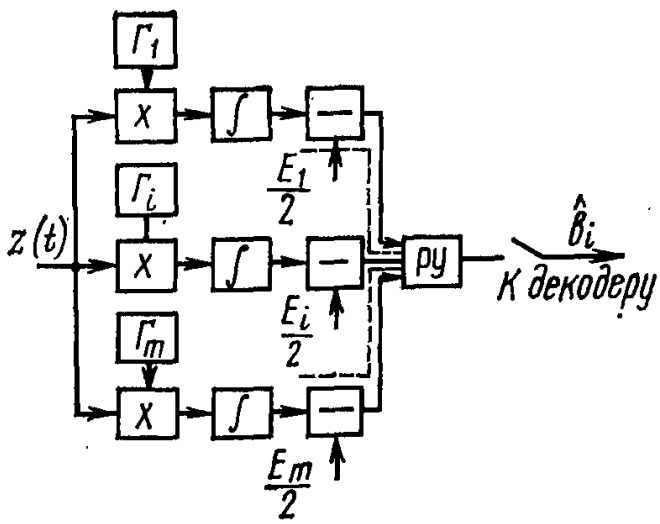


Рис. 4.3. Оптимальный демодулятор при точно известных сигналах

Если сигналы $u_i(t)$ выбраны таким образом, что все их реализации (а следовательно, и все реализации $s_i(t)$) имеют одинаковые энергии (E_i = =const), алгоритм прие-

ма (4.26) (и соответственно его реализация) упрощается (отпадает необходимость в вычитающих устройствах) и принимает вид

$$\max_i \left[\int_0^T z(t)\, s_i(t)\, dt \right]. \qquad (4.29)$$

Из (4.29) видно, что правило решения не изменится, если сигнал $Z(t)$, поступающий на вход демодулятора, умножить на любое число. Поэтому система, в которой все реализации сигнала имеют равную энергию, отличается тем, что оптимальный алгоритм приема в ней не требует знания «масштаба» приходящего сигнала или, другими словами, знания коэффициента передачи k канала. Эта важная особенность обусловила широкое распространение систем сигналов с равной энергией, которые обычно называют *системами с активной паузой*. Особенно важна эта особенность для каналов с замираниями, в которых коэффициент передачи флуктуирует (см. ниже § 4.7).

Следует подчеркнуть, что правильный тактовый синхронизм для выявления границ посылок (съем сигналов на выходе блока РУ в моменты времени, кратные T, и сброс напряжения с интегратора после принятия решения) является непременным условием практической реализации рассмотренных алгоритмов по схеме рис. 4.3.

Для наиболее распространенной двоичной системы из неравенств (4.26) остается лишь одно, и алгоритм приема можно представить в более простом виде:

$$\int_0^T z(t)\, s_1(t)\, dt - 0{,}5\, E_1 > \int_0^T z(t)\, s_2(t)\, dt - 0{,}5\, E_2, \text{ или } \int_0^T z(t)\, s_\Delta(t)\, dt > \lambda, \qquad (4.30)$$

где $s_\Delta(t) = s_1(t) - s_2(t)$ *разностный сигнал*; $\lambda = 0{,}5(E_1 - E_2)$ — *пороговый уровень*. Для системы с активной паузой $\lambda = 0$, что значительно облегчает реализацию оптимальной схемы.

При выполнении неравенства (4.30) регистрируется символ 1, в противном случае — 0. Для реализации (4.30) в схеме рис. 4.3 требуется лишь одна ветвь.

На рис. 4.4 показана схема, реализующая алгоритм (4.30) для двоичной системы передачи однополярными импульсами (с пассивной паузой): $s_1(t) = a$,

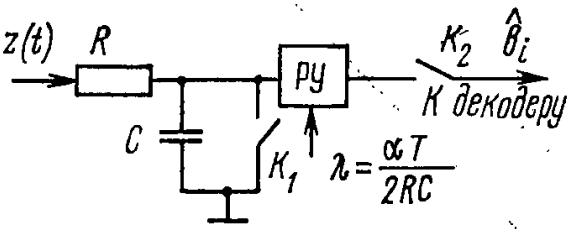


Рис. 4.4. Реализация оптимального приема двоичных прямоугольных видеоимпульсов

$s_2(t) = 0$. При этих сигналах $s_\Delta(t) = s_1(t) = a$, $E_1 = a^2T$, $E_2 = 0$, $\lambda = a^2T/2$ и правило (4.30) примет следующий вид:

$$\int_0^T z(t)\, dt > aT/2. \qquad (4.30а)$$

Интегрирование в схеме рис. 4.4 осуществляется с достаточной точностью цепью RC при условии, что $RC \gg T$. При этом на конденсаторе C напряжение в момент T равно $\frac{1}{RC}\int_0^T z(t)\,dt$. Следовательно, правило (4.30а) сводится к тому, что это напряжение должно превысить пороговый уровень $\lambda = aT/2RC$, который и вводится в РУ. При выполнении этого неравенства в РУ записывается 1, при невыполнении — 0. После этой записи (происходящей при замыкании ключа K_2) необходимо произвести сброс напряжения с интегратора, чтобы можно было принимать следующий элемент сигнала. Сброс осуществляется замыканием ключа K_1, разряжающего конденсатор.

Эта же схема, с небольшой модификацией, может использоваться для демодуляции в двоичной системе передачи двухполярными импульсами (с активной паузой): $s_1(t) = a$, $s_2(t) = -a$. При этом $s_\Delta(t) = 2a$, $E_1 = E_2$ и, следовательно, $\lambda = 0$. В этом случае правило (4.30) после сокращения принимает вид

$$\int_0^T z(t)\,dt > 0. \qquad (4.30б)$$

Его реализует схема рис. 4.4, если пороговый уровень λ положить равным нулю. При этом РУ превращается в *дискриминатор полярности*, выдающий символ 1, когда на его входе напряжение положительно, и 0 — в противном случае.

Рассмотренные две системы используются в простейших устройствах проводной связи. В радиоканалах, а также в современных кабельных каналах используются высокочастотные сигналы. Наиболее простыми двоичными системами с гармоническими сигналами являются системы с амплитудной (АМ), фазовой (ФМ) и частотной (ЧМ) манипуляцией.

В двоичной АМ $s_1(t) = a\cos(\omega_0 t + \varphi)$, $s_2(t) = 0$. Все входящие сюда постоянные (a, ω_0, φ) в этом параграфе полагаем известными. Поскольку здесь $s_\Delta(t) = s_1(t)$, $E_1 = a^2T/2$ и $E_2 = 0$, правило (4.30) запишется так:

$$\int_0^T z(t)\cos(\omega_0 t + \varphi)\,dt > aT/4.$$

Оно реализуется схемой рис. 4.5, которая отличается от рис. 4.4 блоком перемножения приходящего сигнала с опорным сигналом $\cos(\omega_0 t + \varphi)$. Пороговый уровень λ в этом случае равен $aT/(4RC)$.

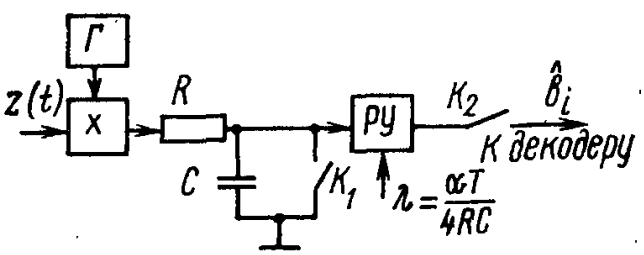


Рис. 4.5. Реализация оптимального приема в двоичной системе АМ, ФМ при точно известном сигнале

При двоичной ФМ системе $s_1(t) = a\cos(\omega_0 t + \varphi)$, $s_2(t) = a\cos(\omega_0 t + \varphi + \pi) = -s_1(t)$. Это — система с активной паузой, и поэтому в (4.30) $\lambda = 0$. Легко убедиться, что правило решения сводится при этом к следующему: $\int_0^T z(t)\cos(\omega_0 t + \varphi)\,dt > 0$ — и реали-

зуется той же схемой рис. 4.5 при $\lambda=0$. В этом случае РУ играет роль дискриминатора полярностей.

Рассмотрим вкратце случай, когда гауссовский шум в канале не белый и не квазибелый, а «окрашенный», т. е. имеет неравномерную спектральную плотность мощности $G(f)$ в полосе спектра сигнала. Пропустим приходящую на вход демодулятора сумму сигнала и шума через фильтр с передаточной функцией $k(i2\pi f)$ такой, чтобы в полосе спектра сигнала произведение $G(f)|k(i2\pi f)|^2$ было постоянной величиной N_0. Из всех возможных фильтров с $k(i2\pi f)$, удовлетворяющих этому условию и различающихся только фазо-частотной характеристикой, выберем минимальио фазовый, который является обратимым. Очевидно, что на выходе фильтра шум окажется квазибелым: $G_{вых}(f)=N_0$. Поэтому такой фильтр называется обеляющим.

Сигнал $s_i(t)$ после прохождения через обеляющий фильтр превратится в некоторый другой сигнал, который обозначим $s'_i(t)$. Вид его можно определить, зная $s_i(t)$ и $k(i2\pi f)$. Если теперь подать колебания с выхода обеляющего фильтра на демодулятор, являющийся оптимальным для приема сигналов $s'_i(t)$ $(i=1,\ldots,m)$, то получим схему рис. 4.6*а*. Дока-

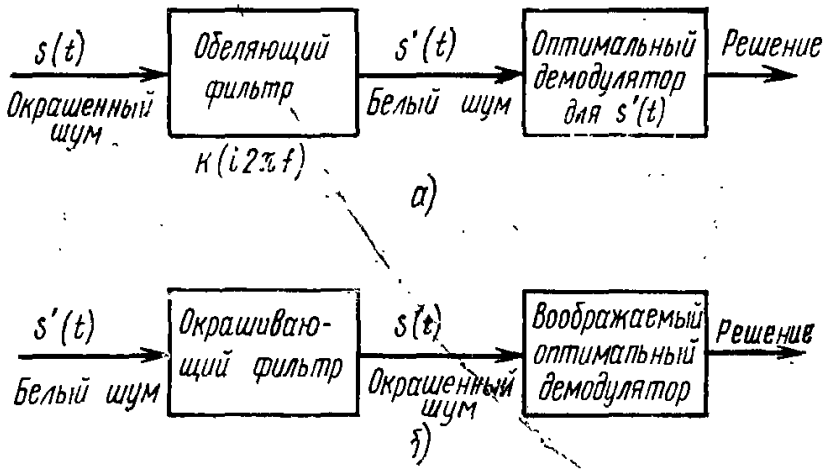


Рис. 4.6. К доказательству оптимальности демодулятора с обеляющим фильтром

жем, что эта схема является оптимальиой для сигналов $s_i(t)$ при окрашенном шуме.

Предположим, что это неверно, т. е. что существует некоторый демодулятор, обеспечивающий меньшую вероятность ошибки, чем демодулятор рис. 4.6*а*, если на тот и другой поступают сигналы $s_i(t)$ на фоне окрашенного шума. (Предполагается, что априорные вероятности всех сигналов одинаковы.) Подключим к входу этого воображаемого оптимального демодулятора фильтр, обратный обеляющему, с передаточной функцией $k^{-1}(i2\pi f)$ (рис. 4.6*б*) и будем подавать на вход этого фильтра сигналы $s'_i(t)$ на фоне белого шума со спектральной плотностью N_0. Легко видеть, что на выходе фильтра будут сигналы $s_i(t)$, а шум будет окрашенным, со спектральной плотностью $G(f)$, т. е. на вход воображаемого оптимального демодулятора будут поступать именно те сигналы и тот шум, на которые он рассчитан. Таким образом, схема рис. 4.6*б* представляет собой демодулятор для сигналов $s'_i(t)$ на фоне белого шума, в котором вероятность ошибок меньше, чем в оптимальном демодуляторе, подключеином к выходу обеляющего фнльтра на рис. 4.6*а*. Это противоречие и доказывает, что не может существовать демодулятор для сигналов $s'_i(t)$ на фоне окрашенного шума лучший, чем на рис. 4.6*а*.

Заметим, что при реализации такого демодулятора с обеляющим фильтром возникают трудности, связанные с тем, что сигналы $s_i(t)$ при прохождении через фильтр, как правило, растягиваются и возникает взаимное наложеине элементов сигнала $s'_i(t)$. Существует ряд путей преодоления этой трудности, однако подробиый анализ их выходит за пределы курса ТПС.

Следует обратить внимание на то, что в схеме рис. 4.5 опорный сигнал должен иметь те же начальные фазы, что и ожидаемые приходящие сигналы или, другими словами, должен быть *когерентным* с приходящими сигналами. Это требование обычно затрудняет реализацию демодулятора и требует введения в него помимо указанных на рис. 4.5 блоков дополнительных устройств, предназначенных для регулировки фаз опорных сигналов.

Все методы приема, для реализации которых необходимо точное априорное знание начальных фаз приходящих сигналов, называются *когерентными*. В тех случаях, когда сведения о начальных фазах ожидаемых сигналов извлекаются из самого принимаемого сигнала (например, если фаза флуктуирует, но настолько медленно, что может быть предсказана по предыдущим элементам сигнала), прием называют *квазикогерентным*. Если же сведения о начальных фазах приходящих сигналов отсутствуют или по каким-либо соображениям не используются, то прием называют *некогерентным* (см. ниже § 4.6).

4.4. РЕАЛИЗАЦИЯ АЛГОРИТМА ОПТИМАЛЬНОГО ПРИЕМА НА ОСНОВЕ СОГЛАСОВАННЫХ ФИЛЬТРОВ

Скалярное произведение (4.28) можно вычислить не только с помощью активного фильтра (коррелятора), описанного в предыдущем параграфе, но и с помощью пассивного линейного фильтра с постоянными параметрами. Если на вход фильтра подать принимаемый сигнал $z(t)$, то напряжение на выходе фильтра $y(t) = \int_0^t g(\tau) z(t-\tau)\, d\tau$; где $g(\tau)$ — импульсная реакция фильтра. Выберем ее такой, чтобы в момент $t=T$ получить значение $y(T)$, совпадающее со скалярным произведением (4.28). Легко видеть, что это будет выполнено, если $g(\tau)=s_i(T-\tau)$. Действительно, при этом $y(T) = \int_0^T s_i(T-\tau) z(T-\tau)\, d\tau = \int_0^T s_i(t) z(t)\, dt = (\mathbf{z}, \mathbf{s}_i)$.

Такой фильтр называется *согласованным* с сигналом $s_i(t)$. В более общем смысле фильтром, согласованным с сигналом $s(t)$, называют линейный фильтр с постоянными параметрами и импульсной реакцией:

$$g(\tau) = a s(t_0 - \tau), \tag{4.31}$$

где a, t_0 — постоянные. Функция $g(t)$ оказывается зеркальным отображением $s(t)$ относительно оси, проведенной через точку $t_0/2$ (рис. 4.7). Для физической реализуемости фильтра необходимо и достаточно, чтобы $g(\tau)=0$ при $\tau<0$. В частности, для финитного сигнала $s(t)$, поступающего на вход фильтра в момент $t=0$ и заканчивающегося в момент T, условие физической реализуемости согласованного фильтра заведомо выполняется, если постоянная t_0 (называемая моментом отсчета) удовлетворяет условию

$$t_0 \geqslant T. \tag{4.32}$$

Действительно, при этом $(t_0-\tau)>T$ и $s(t_0-\tau)=0$, если $\tau<0$. Реакция согласованного фильтра на финитный сигнал $s(t)$ длительностью T существует лишь на финитном интервале протяженностью $2T$.

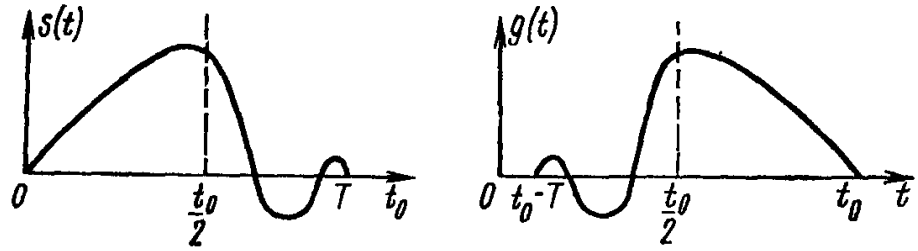


Рис. 4.7. Сигнал $s(t)$ и импульсная реакция $g(t)$ линейного фильтра, согласованного с этим сигналом

Передаточная функция согласованного фильтра с импульсной реакцией (4.31) определяется преобразованием Фурье

$$k(i\omega)_{СФ}=\int_{-\infty}^{\infty} g(t)\exp(-i\omega t)\,dt=a\int_{-\infty}^{\infty} s(t_0-t)\exp(-i\omega t)\,dt=$$

$$=a\int_{-\infty}^{\infty} s(\tau)\exp[-i\omega(t_0-\tau)]\,d\tau=aS^*(i\omega)\exp(-i\omega t_0), \qquad (4.33)$$

где $S^*(i\omega)$ — функция, комплексно-сопряженная со спектральной плотностью сигнала $s(t)$. Следовательно, с точностью до коэффициента a амплитудно-частотная характеристика согласованного фильтра определяется амплитудным спектром сигнала $s(t)$ (т. е. фильтр хорошо пропускает те частоты, которые дают больший вклад в энергию сигнала), а его фазо-частотная характеристика (без учета слагаемого — ωt_0, определяемого задержкой t_0) обратна по знаку фазовой характеристике сигнала $s(t)$. В дальнейшем будем полагать $t_0=T$, поскольку увеличение t_0 не дает никаких преимуществ и только удлиняет задержку в демодуляторе.

Если фильтр согласован с сигналом $s(t)$, а на его входе действует колебание $z(t)$ (в частности, смесь $z(t)=s(t)+n(t)$), то отклик фильтра в соответствии с интегралом Дюамеля в момент отсчета T

$$y(T)=a\int_0^T z(T-\tau)\,s(T-\tau)\,d\tau=a\int_0^T z(x)\,s(x)\,dx \qquad (4.34)$$

Это есть не что иное, как скалярное произведение $a(\mathbf{z}, \mathbf{s})$ в пространстве финитных функций.

Таким образом, в момент времени T напряжение на выходе согласованного фильтра пропорционально сигналу на выходе интегратора активного фильтра в схеме рис. 4.4. Поэтому демодулятор, реализующий алгоритм (4.26), может быть выполнен и на базе согласованных фильтров. Структурная схема такого демоду-

лятора показана на рис. 4.8, где $СФ_i$ — фильтр, согласованный с сигналом $s_i(t)$.

Если на вход фильтра подан тот сигнал, с которым он согласован, то сигнальная составляющая на выходе согласованного фильтра

$$\dot{y}_c(t) = a\int_0^t s(t-\tau)\,s(t_0-\tau)\,d\tau = a\,B_s(t_0-t), \qquad (4.35)$$

где $B_s(t_0-t)$ — временна́я функция корреляции сигнала. Следовательно, формы полезного сигнала на входе и выходе согласованно-

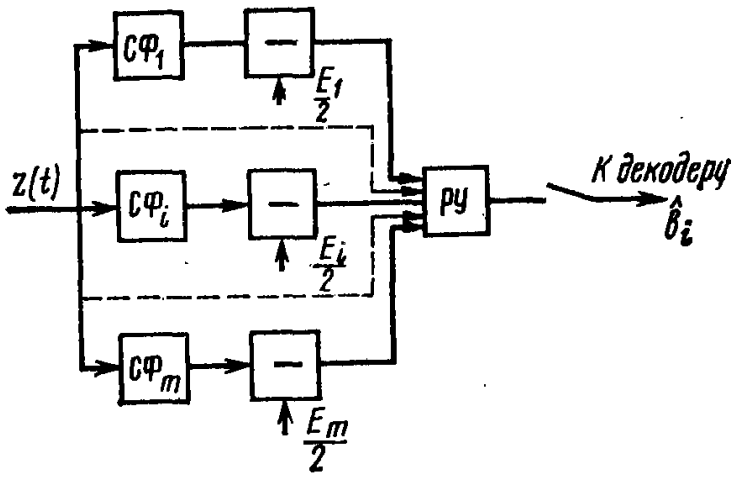


Рис. 4.8. Оптимальный демодулятор на основе согласованных фильтров

го фильтра, как правило, существенно отличаются друг от друга. Следует еще раз подчеркнуть, что задачей согласованного фильтра является не восстановление формы сигнала, искаженной шумом, а получение одного отсчета, по которому можно судить о присутствии или отсутствии на входе фильтра сигнала известной формы.

Отметим одно важное свойство согласованного фильтра, которое иногда рассматривается как его определение. Будем подавать сумму детерминированного сигнала и белого шума $Z(t)=s(t)+N(t)$ на вход различных линейных цепей с постоянными параметрами и измерять в момент $t=t_0$ отношение мгновенной мощности сигнальной составляющей к средней мощности шума на выходе цепи. Докажем, что это отношение $\rho_{пик}$ максимально, если цепь является согласованным фильтром.

Пусть $S(i\omega)$ — спектр входного сигнала, а $k(i\omega)$ — передаточная функция некоторой линейной цепи. Тогда спектр сигнальной составляющей $y_c(t)$ на выходе цепи равен $S(i\omega)k(i\omega)$. С помощью обратного преобразования Фурье найдем значение в момент $t=t_0$:

$$y_c(t_0) = \int_{-\infty}^{\infty} S(i\omega)\,k(i\omega)\exp(i\omega t_0)\,dt,\quad f=\omega/2\pi. \qquad (4.36)$$

Мощность шума на выходе цепи $P_ш = (N_0/2)\int_{-\infty}^{\infty} |k(i\omega)|^2 df$. Искомое отношение

$$\rho_{пик} = \frac{|y_c(t_0)|^2}{P_ш} = \frac{\left|\int_{-\infty}^{\infty} S(i\omega)k(i\omega)\exp(i\omega t_0)\,df\right|^2}{\frac{N_0}{2}\int_{-\infty}^{\infty}|k(i\omega)|^2 df}. \qquad (4.37)$$

Согласно неравенству Буняковского — Шварца (2.93), (2.103) для любых комплексных функций $A(x)$ и $B(x)$ при $b>a$ $\left|\int_a^b A^*(x)B(x)dx\right|^2 \leqslant \int_a^b |A(x)|^2 dx \int_a^b |B(x)|^2 dx$, причем равенство имеет место только при

$$B(x) = \alpha A(x), \qquad (4.38)$$

где α — произвольная постоянная.

Применим неравенство Буняковского — Шварца к числителю (4.37). Если положить

$$A(\omega) = S^*(i\omega)\exp(-i\omega t_0);\ B(\omega) = k(i\omega);\ A^*(\omega) = S(i\omega)\exp(i\omega t_0),$$

то получим

$$\rho_{пик} \leqslant \frac{\int_{-\infty}^{\infty}|S(i\omega)\exp(i\omega t_0)|^2 df \int_{-\infty}^{\infty}|k(i\omega)|^2 df}{\frac{N_0}{2}\int_{-\infty}^{\infty}|k(i\omega)|^2 df} = \frac{2}{N_0}\int_{-\infty}^{\infty}|S(i\omega)|^2 df =$$

$$= \frac{2E}{N_0} = 2h^2, \qquad (4.39)$$

где $h^2 = E/N_0$ — отношение энергии элемента сигнала E на входе фильтра к спектральной плотности N_0 белого шума.

Знак равенства согласно (4.38) и (4.39) имеет место тогда, когда передаточная функция удовлетворяет (4.33), т. е. для согласованного фильтра, что и требовалось доказать.

Рассмотрим вкратце возможности реализации согласованных фильтров. Согласованный фильтр для финитного сигнала произвольного вида $s(t)$ можно, в принципе, построить на основе неискажающей длинной линии, обеспечивающей задержку сигнала на время T, с бесконечной плотностью отводов. Практически можно брать отводы в дискретных точках с разносом $\Delta t = 0{,}5F$, где F — эффективная ширина спектра сигнала.

Действительно, с помощью схемы, показанной на рис. 4.9, можно с хорошей точностью синтезировать любой сигнал $s(t)$, представленный усеченным рядом Котельникова (2.116):

$$s(t) = \sum_{k=0}^{n} a_k \frac{\sin 2\pi F(t - k\Delta t)}{2\pi F(t - k\Delta t)},$$

где $a_k = s(k\Delta t)$; $\Delta t = 1/(2F)$; F — ширина спектра сигнала. Как было показано в § 2.7 (с. 66), такой сигнал можно получить на выходе идеального ФНЧ с полосой пропускания F, подавая на его вход последовательность δ-импульсов с весами a_k через интервалы времени Δt. Это осуществляется с известным приближением в схеме рис. 4.9. Если на вход A линии в начальный момент пода-

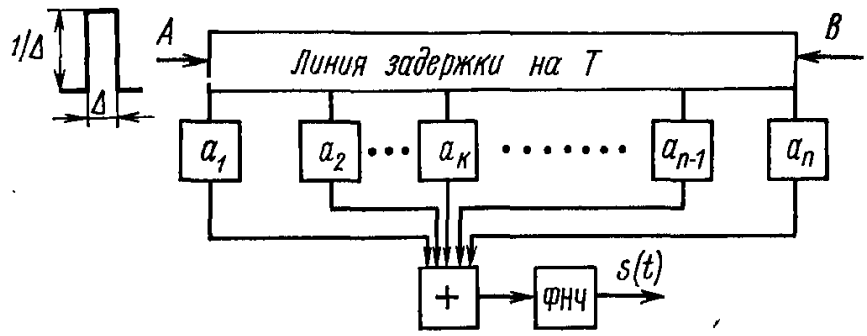


Рис. 4.9. Реализация фильтра, согласованного с произвольным непрерывным сигналом на основе линии задержки с отводами и блоками взвешивания

ется один короткий единичный импульс, аппроксимирующий δ-функцию, то с отводов снимаются такие же импульсы, разнесенные на интервалы Δt, которые, пройдя через взвешивающие блоки a_k, поступают поочередно на вход ФНЧ. Взвешивающие блоки содержат аттенюаторы или усилители с коэффициентом усиления $|a_k|$, а также при отрицательных a_k инверторы.

Таким образом, схема рис. 4.9 представляет собой линейный фильтр с импульсной реакцией $s(t)$. Легко видеть, что если входной импульс подать не в точку A, а в точку B, то будет синтезирован сигнал, представляющий зеркальное отображение $s(t)$. Поэтому та же схема со входом в точке B оказывается фильтром, согласованным с $s(t)$.

Существуют различные другие способы реализации фильтра, точно или приближению согласованного с сигналом заданной формы. Так, например, фильтр, согласованный с прямоугольным видеоимпульсом длительностью T, который по определению (4.31) должен иметь импульсную реакцию также в форме прямоугольного импульса, можно построить по схеме рис. 4.10*а*, содержащей конденсатор, линию задержки на время T, инвертор и сумматор. Если на вход этого фильтра подать единичный импульс (дельта-функцию), то он зарядит конденсатор до некоторого напряжения. Затем, спустя время T, через линию задержки и инвертор такой же входной импульс противоположной полярности поступит на конденсатор и разрядит его. Как следствие на выходе фильтра образуется в качестве импульсной реакции прямоугольный импульс.

Вариант фильтра, согласованного с прямоугольным радиоимпульсом $s(t) = a\cos(\omega_0 t + \varphi_0)$, заданным на интервале $(0, T)$, показан на рис. 4.10*б*. Он состоит из идеального колебательного контура без потерь[1], настроенного на частоту ω_0, и фазовращателя, сдвигающего фазу колебаний контура на $-\omega_0 T - \varphi_0$. Линия задержки на T в схеме и инвертор обеспечивают гашение колебаний фильтра вне интервала $(0, T)$, импульсная же реакция фильтра на этом интервале с учетом фазовращателя $a\cos[\omega_0(t-T)-\varphi_0] = a\cos[\omega_0(T-t)+\varphi_0] = as(T-t)$, что и обеспечивает согласование согласно (4.31).

Другой вариант фильтра с коммутацией параметров (так называемый кинематический фильтр), заменяющий фильтр, согласованный с прямоугольным радиоимпульсом, показан на рис. 4.10*в*. Он содержит высокодобротный колеба-

[1] В реальном фильтре всегда имеются потери, но если затухание контура β достаточно мало ($\beta T \ll 1$), то импульсная реакция контура $ae^{-\beta t}\cos\omega_0 t$ при $t \leqslant T$ практически не отличается от $a\cos\omega_0 t$. Для уменьшения β применяют электромеханические фильтры (кварцевые и др.), а также используют схемы с положительной обратной связью.

тельный контур, обеспечивающий на частоте ω_0 фазовый сдвиг — $\omega_0 T$—φ_0, и устройство, позволяющее гасить колебания в контуре в моменты, кратные T, после снятия отсчета (на схеме это осуществляется путем закорачивания емкостной и разрыва индуктивной ветвей) для освобождения контура от накопленной энергии и подготовки его к приему следующих элементов сигнала.

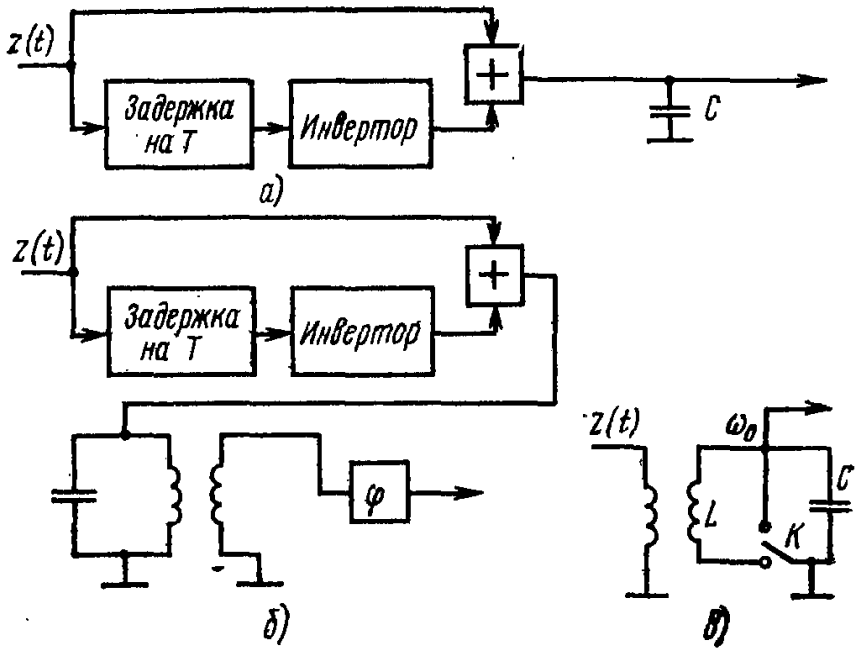


Рис. 4.10. Реализация согласованных факторов:
а) для прямоугольного видеоимпульса; *б)* для прямоугольного радиоимпульса; *в)* то же (фильтр с коммутацией параметров)

На рис. 4.11 и 4.12 изображены оптимальные схемы приема двоичных сигналов АМ, ФМ (с частотой заполнения ω_0) и ЧМ (с частотами ω_1 и ω_2), построенные на основе коммутируемого фильтра.

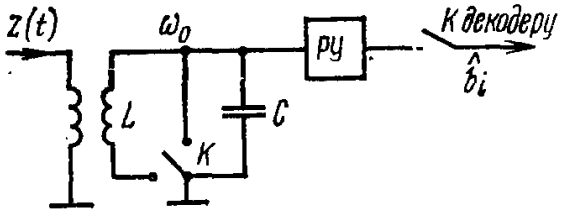


Рис. 4.11. Оптимальный демодулятор в системе АМ или ФМ на базе коммутируемого фильтра

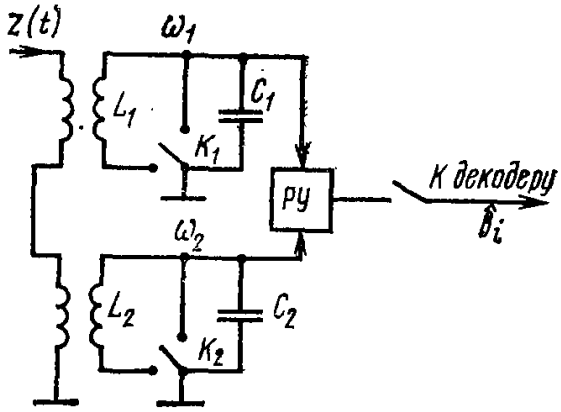


Рис. 4.12. Оптимальный демодулятор в системе ЧМ на базе коммутируемых фильтров

Схема с согласованными фильтрами на первый взгляд кажется проще схемы с активными фильтрами, поскольку в ней нет опорных генераторов и не возникает проблемы обеспечения их когерентности (согласования по фазе с приходящим сигналом). Однако и в схеме с согласованными фильтрами имеются свои реа-

лизационные трудности. В этом можно убедиться, сравнив, например, эпюры напряжений (без учета помех в канале) на выходе фильтра (рис. 4.13*б*), согласованного с прямоугольным радиоимпульсом (рис. 4. 13*а*), и на выходе интегратора активного фильтра (рис. 4.13*в*). Отметим, что всюду, за исключением точки $t=T$, напряжения на выходах обоих фильтров отличаются друг от друга. Из рисунков видно, что допустимая неточность во времени Δt снятия отсчета максимума сигнала на выходе активного фильтра значительно больше, чем при снятии отсчета максимума сигнала на выходе согласованного фильтра. При активном фильтре достаточно потребовать, чтобы неточность взятия отсчета была мала по сравнению с тактовым интервалом T, а при согласованном фильтре—по сравнению с периодом высокочастотного заполнения радиоимпульса (так называемый *когерентный отсчет*). Трудность обеспечения когерентного отсчета в согласованном фильтре вполне соизмерима с трудностью реализации когерентных опорных генераторов в активном фильтре.

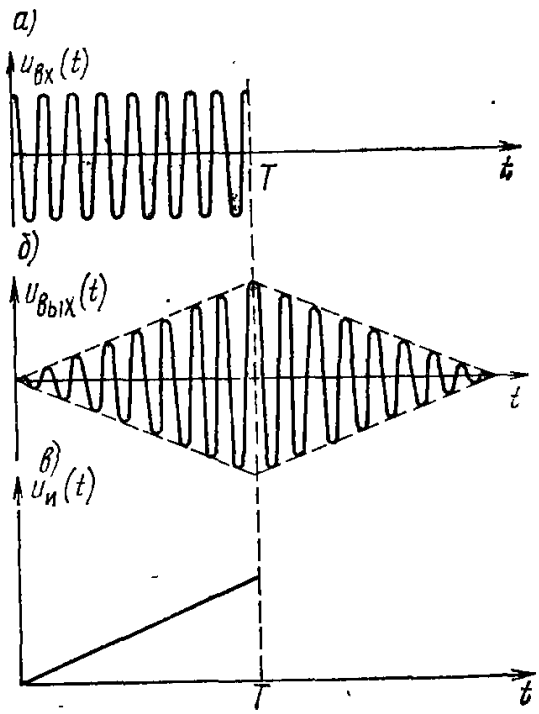


Рис. 4.13. Сигналы на выходе согласованного фильтра и корреляционной схемы при подаче на вход прямоугольного радиоимпульса:
а) импульс на входе; *б)* импульс на выходе согласованного фильтра; *в)* напряжение на выходе интегратора активного фильтра

Отметим еще одно свойство согласованного фильтра, вытекающее из его линейности.

Если повернуть фазы всех составляющих входного сигнала на $-90°$ в области $\omega>0$ и на $90°$ в области $\omega<0$, т. е. заменить входной сигнал $z(t)$ сопряженным с ним по Гильберту $\tilde{z}(t)$, то и фазы выходного сигнала $y(t)$ повернутся таким же образом. Следовательно, сигнал, сопряженный с выходным напряжением (4.34) согласованного фильтра при подаче на вход сигнала $\tilde{z}(t)$,

$$\tilde{y}(t)=a\int_0^t \tilde{z}(t-\tau)\,s(t_0-\tau)\,d\tau=a\int_0^t z(t-\tau)\,\tilde{s}(t_0-\tau)\,d\tau. \qquad (4.40)$$

Другими словами, безразлично, где осуществлять поворот фаз, до фильтра или после него. Исходя из этого, можно определить огибающую $r(t)$ на выходе фильтра, согласованного с сигналом $s(t)$ $(0\leqslant t\leqslant T)$, при подаче на вход колебания $z(t)$:

$$r(t)=a\sqrt{\left[\int_0^t z(x)\,s(T-t+x)\,dx\right]^2+\left[\int_0^t z(x)\,\tilde{s}(T-t+x)\,dx\right]^2}. \qquad (4.41)$$

В момент времени $t=T$

$$r(T)=a\sqrt{\left[\int_0^T z(x)\,s(x)\,dx\right]^2+\left[\int_0^T z(x)\,\tilde{s}(x)\,dx\right]^2}. \qquad (4.42)$$

Легко показать, что величина $r(T)$ инвариантна к изменению фазы сигнала. Пусть, например, $s_\theta(t)=s(t)\cos\theta+\tilde{s}(t)\sin\theta$. Сопряженный с ним сигнал $\tilde{s}_\theta(t)=\tilde{s}(t)\cos\theta-s(t)\sin\theta$.

Подставляя эти значения в подкоренное выражение в формуле (4.41), получим после простых преобразований $[\int_0^t z(x)s_\theta(x)dx]^2+[\int_0^t z(x)\tilde{s}_\theta(x)dx]^2=$

$=[\int_0^t z(x)s(x)dx]^2+[\int_0^t z(x)\tilde{s}(x)dx]^2$, независимо от θ.

Следовательно, если мы интересуемся лишь огибающей сигнала на выходе согласованного фильтра, то согласование может быть выполнено с точностью до фазы, что во многих случаях упрощает реализацию схемы.

Из (4.42) следует, что если на вход фильтра, согласованного с сигналом $s(t)$, подать сигнал $z(t)$, то огибающая на выходе фильтра в момент отсчета T равна нулю тогда и только тогда, когда сигналы $s(t)$ и $z(t)$ ортогональны на интервале $(0, T)$ в усиленном смысле (см. с. 64). Легко убедиться, что если $s(t)$ и $z(t)$ ортогональны, но не в усиленном смысле, то в момент отсчета мгновенное напряжение сигнала на выходе фильтра равно нулю, однако огибающая отлична от нуля.

В технике связи для фильтрации сигнала на фоне шума часто вместо согласованных используют фильтры, характеристики которых лишь частично согласованы с характеристиками сигнала. Такие фильтры называют квазиоптимальными.

Так, в практике радиоприема используются так называемые квазиоптимальные линейные фильтры, форма частотных характеристик которых заранее задана и максимум параметра $\rho_{пик}$ обеспечивается лишь соответствующим подбором ширины полосы пропускания фильтра. Квазиоптимальный фильтр такого типа впервые исследовался В. И. Сифоровым, который рассматривал прохождение одиночного радиоимпульса с прямоугольной огибающей через идеальный полосовой фильтр с полосой пропускания Δf на фоне квазибелого шума.

Воспользовавшись формулой (4.37), В. И. Сифоров показал, что при $\Delta f\approx 1{,}37/T$ отношение $\rho_{пик}$ достигает максимума, равного

$$\rho_{пик}\approx 1{,}64\,h^2. \qquad (4.43)$$

Сравнив (4.39) и (4.43), можно видеть, что при приеме одиночного импульса энергетический выигрыш оптимального фильтра по сравнению с квазиоптимальным невелик (не превышает 1 дБ).

Таким образом, при приеме одиночных радиоимпульсов вполне допустимо ограничиться квазиоптимальной фильтрацией. Положение, однако, существенно меняется, если надлежит принимать информационные импульсы, следующие друг за другом с таким интервалом, на котором переходные процессы в квазиоптимальном фильтре не успевают затухнуть. В этих условиях качество приема с квазиоптимальной фильтрацией резко падает, в то время как при использовании оптимального согласованного фильтра качество остается прежним, так как сигнал на его выходе концентрируется на ограниченном временно́м интервале и к моменту отсчета для одного импульса реакция на все предыдущие импульсы равна нулю (см., например, рис. 4.13*б*).

4.5. ПОТЕНЦИАЛЬНАЯ ПОМЕХОУСТОЙЧИВОСТЬ ПРИ ТОЧНО ИЗВЕСТНОМ АНСАМБЛЕ СИГНАЛОВ

Определим потенциальную помехоустойчивость для двоичной системы с аддитивным белым шумом, когда при приеме точно известны оба ожидаемых сигнала: $s_1(t)$ и $s_2(t)$, полагая, что априорные вероятности этих сигналов одинаковы. Приходящий сигнал $Z(t)$ является случайным, так как, во-первых, он содержит случайную помеху $N(t)$, во-вторых, заранее неизвестна реализация передаваемого сигнала.

В этом случае согласно (4.30) алгоритм оптимального приема можно записать в виде

$$\int_0^T Z(t)\,[s_1(t)-s_2(t)]\,dt > 0{,}5\,(E_1-E_2);\quad E_i=\int_0^T s_i^2(t)\,dt. \tag{4.44}$$

При выполнении неравенства (4.44) оптимальный приемник регистрирует символ 1, соответствующий сигналу $s_1(t)$, в противном случае — символ 0, соответствующий сигналу $s_2(t)$. Если действительно передается символ 1, то $Z(t)=s_1(t)+N(t)$. При этом вероятность ошибки $p(0|1)$ определится вероятностью того, что неравенство (4.44) не выполнено, т. е. вероятностью выполнения неравенства

$$\int_0^T s_1(t)\,[s_1(t)-s_2(t)]\,dt+\int_0^T N(t)\,[s_1(t)-s_2(t)]\,dt \leqslant 0{,}5\int_0^T [\,s_1^2(t)-s_2^2(t)]\,dt, \tag{4.45}$$

которое легко привести к следующему виду:

$$\int_0^T N(t)\,[s_1(t)-s_2(t)]\,dt \leqslant -0{,}5\int_0^T [s_1(t)-s_2(t)]^2\,dt. \tag{4.46}$$

Аналогичное соотношение получится, если предположить, что передается символ 0. Следовательно, в обоих случаях вероятности ошибки равны: $p(0|1)=p(1|0)=p$ и сформированный модемом двоичный дискретный канал симметричен.

Запишем (4.46) в виде

$$\xi \leqslant -0{,}5\,E_э, \tag{4.47}$$

где $\xi=\int_0^T N(t)s_\Delta(t)dt$; $E_э=\int_0^T s^2{}_\Delta(t)dt$; $s_\Delta(t)=s_1(t)-s_2(t)$.

Если $N(t)$ — нормальный стационарный белый шум с нулевым средним и односторонней спектральной плотностью мощности N_0, то ξ — нормально распределенная величина, так как она определяется линейной операцией над нормальным же случайным процессом.

Ее математическое ожидание

$$\overline{\xi}=\int_0^T \overline{N(t)}\,s_\Delta(t)\,dt=0, \tag{4.48}$$

а дисперсия $D(\xi)=\overline{\xi^2}=\overline{[\int_0^T N(t)s_\Delta(t)dt]^2}=\int_0^T\int_0^T \overline{N(t_1)N(t_2)}\,s_\Delta(t_1)\times$
$\times s_\Delta(t_2)dt_1dt_2$. Учитывая выражение (2.49) для ФК белого шума $\overline{N(t_1)N(t_2)}$, получим

$$D(\xi)=\int_0^T\int_0^T \frac{N_0}{2}\,\delta(t_2-t_1)\,s_\Delta(t_1)\,s_\Delta(t_2)\,dt_1dt_2=0{,}5\,N_0\int_0^T s_\Delta^2(t)\,dt=$$

$$=0{,}5\,N_0E_э. \qquad (4.49)$$

Поэтому вероятность выполнения неравенства (4.47), т. е. вероятность ошибки,

$$p=\int_{-\infty}^{-0{,}5E_э} w(\xi)\,d\xi=\int_{-\infty}^{-0{,}5E_э}\frac{1}{\sqrt{2\pi D(\xi)}}\exp\left(-\frac{\xi^2}{2D(\xi)}\right)d\xi=$$

$$=\int_\alpha^\infty\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\eta^2}{2}\right)d\eta, \qquad (4.50)$$

где произведена замена переменной $\eta=-\xi/\sqrt{D(\xi)}$ и введено обозначение $\alpha=0{,}5E_э/\sqrt{D(\xi)}=0{,}5E_э/\sqrt{0{,}5N_0E_э}=\sqrt{E_э/2N_0}$.

Функция $\Phi(x)=\frac{2}{\sqrt{2\pi}}\int_0^x e^{-\eta^2/2}d\eta$ табулирована и называется функцией Крампа. Учитывая, что $\Phi(\infty)=1$, можно записать (4.50) в следующем виде:

$$p=0{,}5\left[1-\Phi\left(\sqrt{\frac{E_э}{2N_0}}\right)\right]. \qquad (4.51)$$

При заданной интенсивности помехи N_0 потенциальная помехоустойчивость двоичной системы зависит только от так называемой эквивалентной энергии сигналов:

$$E_э=\int_0^T[s_1(t)-s_2(t)]^2\,dt, \qquad (4.52)$$

которая равна квадрату расстояния между сигнальными точками в пространстве Гильберта. Помехоустойчивость выше (вероятность ошибки меньше) у той системы, у которой больше эквивалентная энергия используемых сигналов.

Соотношение (4.52) позволяет осуществлять оптимальный выбор сигналов $s_1(t)$ и $s_2(t)$ или соответственно $u_1(t)$ и $u_2(t)$, обеспечивающих максимально возможную помехоустойчивость при заданной энергии сигналов E. В самом деле, для такой оптимальной системы величина $E_э$ должна быть максимальной при условии, что

$$E_1=\int_0^T s_1^2(t)\,dt\leqslant E;\quad E_2=\int_0^T s_2^2(t)\,dt<E. \qquad (4.53)$$

Можно написать $E_э = 2E_1 + 2E_2 - \int_0^T [s_1(t) + s_2(t)]^2 dt$. Для получения максимума этого выражения нужно сделать E_1 и E_2 возможно бо́льшими, а интеграл в правой части — как можно меньшим. Максимально возможные значения E_1 и E_2 получатся, если, учитывая условия (4.53), положить

$$E_1 = E_2 = E. \tag{4.54}$$

Интеграл $\int_0^T [s_1(t) + s_2(t)]^2 dt$ принимает только неотрицательные значения, поэтому его минимум равен нулю и достигается при условии

$$s_1(t) = -s_2(t), \tag{4.55}$$

которое не противоречит условию (4.54). Таким образом, в двоичном канале с постоянными параметрами и аддитивной флуктуационной помехой оптимальной оказывается система с противоположными сигналами (4.55). Этому условию удовлетворяют, например, двухполярные импульсы (см. с. 129), сигналы двоичной фазовой модуляции (ФМ), если разность фаз сигналов $\Delta\varphi = \pi$, и т. д. Для всех таких систем $E_э = 4E$ и вероятность ошибки

$$p = 0{,}5\left[1 - \Phi\sqrt{\frac{2E}{N_0}}\right] = 0{,}5\left[1 - \Phi\left(\sqrt{2}h\right)\right], \tag{4.56}$$

где $h^2 = E/N_0$ — отношение энергии сигнала на входе демодулятора к спектральной плотности флуктуационной помехи.

Для системы с активной паузой и ортогональными сигналами (например, при известных условиях для системы двоичной ЧМ), когда $\int_0^T s_1(t)s_2(t)dt = 0$, $E_э = 2E$ и минимальная вероятность ошибки

$$p = 0{,}5\left[1 - \Phi\sqrt{\frac{E}{N_0}}\right] = 0{,}5\left[1 - \Phi(h)\right]. \tag{4.57}$$

Сравнивая (4.56) и (4.57), приходим к выводу, что переход от системы с ортогональными сигналами к системе с оптимальными (противоположными) сигналами позволяет в рассматриваемом канале обеспечить неизменное качество связи (вероятность ошибки) при понижении средней мощности передатчика в 2 раза, т. е. дает энергетический выигрыш в 2 раза (или на 3 дБ).

В двоичной системе с пассивной паузой, полагая $s_2(t) = 0$ и $\int_0^T s_1^2(t)dt = E$, получаем для минимальной вероятность ошибки

$$p = 0{,}5\left[1 - \Phi\left(h\sqrt{2}\right)\right]. \tag{4.58}$$

Отсюда видно, что при переходе от системы АМ к системе ЧМ энергетический выигрыш по максимальной мощности равен 2, а при переходе к системе ФМ — 4. Если же сравнение вести не по пиковой, а по средней мощности, то переход от АМ к ЧМ не дает энергетического выигрыша, поскольку при ЧМ средняя мощность равна максимальной, а при АМ — вдвое меньше максимальной (если s_1 и s_2 передаются с одинаковой вероятностью).

Тем не менее, когда в начале 40-х годов в радиосвязи стали применять ЧМ, помехоустойчивость значительно возросла по сравнению с ранее господствовавшей системой АМ. Это объясняется не увеличением потенциальной помехоустойчивости, которая для обеих систем одинакова, а, главным образом, тем, что оптимальная решающая схема для ЧМ реализуется с довольно большой точностью, а при АМ этому препятствует невозможность обеспечить точное оптимальное значение ненулевого порогового уровня λ (см. рис. 4.5). Поэтому реальная помехоустойчивость при ЧМ близка к потенциальной, а при АМ значительно ниже ее.

Система ФМ (см. с. 129), как и другие системы с противоположными сигналами, обеспечивает максимальную для двоичной системы потенциальную помехоустойчивость. Однако реализация демодулятора для когерентного приема ФМ встречает серьезные трудности. При построении демодулятора с активным фильтром (см. рис. 4.5) возникает проблема поддержания равенства фаз опорного генератора и приходящего сигнала. Если же пытаться его строить на основе согласованного фильтра (например, рис. 4.11), то возникает еще более трудная задача когерентного отсчета.

В практических схемах опорный сигнал $s_1(t)$ формируется из принимаемого колебания, поскольку если его генерировать автономным генератором в месте приема, то необходимое согласование по фазе не может быть обеспечено вследствие неизбежных флуктуаций. Для этого нужно по принимаемому сигналу восстановить немодулированный гармонический сигнал $\cos(\omega_0 t+\varphi)$. При АМ это осуществляется просто, путем выделения фильтром составляющей спектра на несущей частоте. При ФМ этого сделать не удается, так как если элемеиты $s_1(t)$ и $s_2(t)=-s_1(t)$ передаются равновероятно, то спектр сигнала ФМ вообще не содержит составляющей на частоте ω_0. Для ее получения приходится использовать нели-

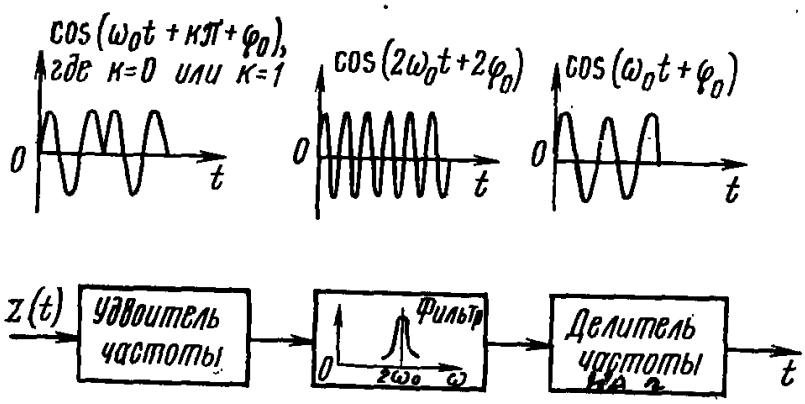


Рис. 4.14. Метод А. А. Пистолькорса для формирования опорного сигнала при ФМ с манипуляцией на π

нейные устройства снятия манипуляции. Это достигается различными схемами, например схемой, предложенной А. А. Пистолькорсом (рис. 4.14). Схема содержит умножитель частоты на 2, выходной сигнал которого через узкополосный фильтр, настроенный на частоту $2\omega_0$, поступает на делитель частоты на 2.

Все схемы формирования опорного сигнала таковы, что вследствие различных неконтролируемых факторов возможны случайные изменения знака опорного сигнала. Это, в частности, относится и к делителю частоты на 2 в схеме А. А. Пистолькорса, поскольку эта операция неоднозначна — фаза выходного сигнала делителя может принять любое из двух значений — $\varphi/2$ или $\varphi/2+\pi$. Это означает, что символы, регистрируемые в дискретной памяти на выходе приемника даже при отсутствии аддитивной помехи в канале ($z(t)=s_i(t)$), после случайного перескока фазы опорного сигнала инвертируются (нули будут записаны как 1, а 1 как 0). Это будет продолжаться до следующего перескока фазы опорного сигнала. Возникает так называемое явление «обратной работы», вследствие которого практическое внедрение систем с фазовой манипуляцией оказалось невозможным.

Эффективный метод устранения этого явления был найден путем перехода к относительным методам модуляции, предложенным Н. Т. Петровичем (см. § 1.5). При относительной фазовой манипуляции (ОФМ) информация содержится не в абсолютном значении фазы элемента сигнала, а в разности фаз двух соседних элементов. Сигналы ОФМ могут приниматься различными методами. Здесь рассмотрим квазикогерентный метод приема сигналов ОФМ (называемый методом сравнения полярностей). Заметим сначала, что систему ОФМ можно рассматривать как обычную систему с фазовой манипуляцией (ФМ), но со специальным перекодированием символов. Это означает, что оптимальный прием сигналов ОФМ можно осуществить, например, схемой рис. 4.5, но с перекодированием принятых символов. Перекодирование выполняется сравнением полярностей напряжений на выходе интегратора для двух соседних элементов, для чего, естественно, требуется задержка выходных символов в ячейке памяти (ЯП) на время T. Такая схема демодулятора показана на рис. 4.15 (без устройства под-

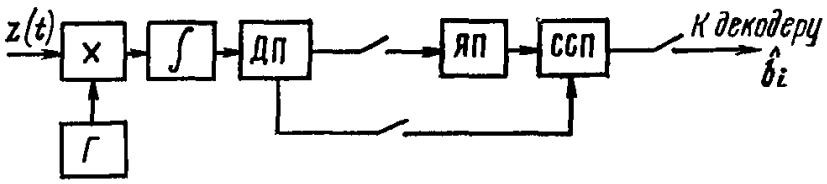


Рис. 4.15. Схема оптимального приема сигналов ОФМ методом сравнения полярностей (когерентный прием)

стройки фазы опорного генератора Γ, которое может быть выполнено, например, по схеме рис. 4.14). Так как ОФМ — система с активной паузой, то пороговый уровень в демодуляторе — нулевой и решающее устройство превращается в дискриминатор полярности (ДП). Полярности соседних элементов сравниваются в схеме сравнения полярностей (ССП), которая представляет собой обычный перемножитель. Символ 1 регистрируется на выходе приемника при совпадении полярностей двух соседних посылок, символ 0 — если эти полярности противоположны. При таком методе

приема ошибка (при отсутствии помехи в канале) возникает в момент перескока фазы опорного сигнала только в одном символе. Последующие же символы регистрируются правильно, т. е. явление «обратной работы» устранено.

Определим вероятность ошибки в системе ОФМ при учете флуктуационной помехи в канале при когерентном приеме. Вероятность $p_{\text{ОФМ}}$ ошибочной регистрации символов в системе ОФМ при приеме по методу сравнения полярностей не совпадает с вероятностью появления ошибок на выходе фазового детектора или, что то же самое, с вероятностью ошибок в системе «классической» фазовой манипуляции (ФМ), определяемой формулой (4.56). Очевидно, что ошибочная регистрация символа при прнеме методом сравнения полярностей возможна в результате одного из двух несовместимых событий: а) знак данного элемента принят ошибочно, а знак предыдущего — верно; б) знак данного элемента принят верио, а предыдущего — ошибочно. Каждое из этих событий имеет вероятность $p_{\text{ФМ}}(1-p_{\text{ФМ}})$. Таким образом,

$$p_{\text{ОФМ}} = 2p_{\text{ФМ}}(1-p_{\text{ФМ}}) = 0{,}5\left[1-\Phi^2\left(\sqrt{2}h\right)\right]. \qquad (4.59)$$

В нормальных условиях эксплуатации, когда требуется $p_{\text{ФМ}} \ll 1$,

$$p_{\text{ОФМ}} \approx 2p_{\text{ФМ}} = 1-\Phi\left(\sqrt{2}h\right).$$

Таким образом, «платой» за устранение обратной работы является удвоение вероятиости ошибки, обусловленной шумом в канале.

Очевидно, что при рассматриваемом методе приема снгиалов ОФМ образующнйся дискретный канал является марковским (см. с. 103). Вероятность ошнбки в нем зависит от того, правильно или ошибочно принят предыдущий символ. Подавляющее большинство ошнбок группнруется по две.

Для недвоичных систем ($m>2$) нахождение вероятности ошибочного приема p_m в общем случае затрудняется, так как теперь приходится анализировать совокупность из $(m-1)$ неравенств. Однако для систем с активной паузой ($E_i=E$) при равновероятных ортогональных сигналах канал симметричен и можно оценить вероятность p_m простым неравенством

$$p_m < (m-1)\,p_2, \qquad (4.60)$$

где p_2 — вероятность ошибки для двоичной системы в том же канале, если используется некоторая пара из m сигналов.

Для доказательства заметим, что решение о том, что передан сигнал $u_i(t)$, принимается, если одновременно выполняются $m-1$ иеравенств (4.26). Если действительно передавался сигнал $u_i(t)$, то ошибка произойдет в том случае, когда хотя бы одно из неравенств (4.26) ие будет выполиено. По формуле Буля (см. сноску на с. 116) вероятность этого не превышает суммы вероятностей иевыполнення каждого из неравенств (4.42). Поскольку прн сделанных предположеннях вероятность невыполнення любого неравенства (4.26) равна p_2, то отсюда следует оценка (4.60). При $p_m \ll 1$ вероятность того, что одновременно два или более иеравеиств в (4.26) не будут выполнены, очень мала. В этом случае, в соответствии с замечанием в сноске на с. 116, иеравенство (4.60) даёт очень плотную оценку для p_m, так что его можно рассматривать как приближеиное равенство.

Из формулы (4.60) не следует делать поспешного вывода о том, что m-ичная система при $m>2$ хуже двоичной. Фиксируя скорость ввода информации в канал в обеих системах $H'(A)=(\log m)/T$, видим, что длительность m-ичного сигнала $T_m=(\log m)/H'(A)$ оказывается равной $T_2 \log m$, где T_2 — длительность двоичного сигнала. От длительности зависит и энергия сигнала $E_m=P_c T_m$, где P_c — мощность принимаемого сигнала. Соответственно при фиксированной величине спектральной плотности шума и параметр h^2 оказывается в m-ичной системе в $\log m$ раз больше, чем в двоичной. Это увеличение h^2 при определенных условиях с избытком компенсирует влияние множителя $(m-1)$ в формуле (4.60).

Пусть, например, используются m взаимно ортогональных сигналов[1]. При $m=2$ вероятность ошибки определяется формулой (4.57). Для удобства дальнейших преобразований заменим эту формулу довольно грубой оценкой, справедливой при любых значениях h:

$$p_2 = 0{,}5\,[1-\Phi(h)] < 0{,}5 \exp(-h^2/2). \tag{4.61}$$

Тогда в m-ичной ортогональной системе

$$p_m < 0{,}5\,(m-1)\exp\left(-h_m^2\,0{,}5\right) = 0{,}5\,(m-1)\exp\left[-0{,}5\,h_2^2 \log m\right] <$$

$$< 0{,}5\,m \exp\left[-0{,}5\,h_2^2 \log m\right],$$

где $h^2_2 = \dfrac{P_c T_2}{N_0} = P_c/[N_0 H'(A)]$; $h^2_m = \dfrac{P_c T_m}{N_0} = h^2_2 \log m$. Имеем

$$p_m < 0{,}5\,m \exp\left[\frac{-0{,}5\,P_c \log m}{N_0 H'(A)}\right] = 0{,}5\,e^{\ln m} \exp\left[\frac{-0{,}5\,P_c \log e \ln m}{N_0 H'(A)}\right] =$$

$$= 0{,}5 \exp\left[\ln m\left(1 - 0{,}5\frac{P_c \log e}{N_0 H'(A)}\right)\right]. \tag{4.62}$$

С увеличением m правая часть неравенства (4.62) стремится к нулю, если $1-0{,}5\dfrac{P_c \log e}{N_0 H'(A)} < 0$ или

$$H'(A) < 0{,}5\frac{P_c}{N_0} \log e. \tag{4.63}$$

Так как мы пользовались очень грубыми оценками [в частности, (4.61)], полученное условие (4.63) является достаточным, но не необходимым для того, чтобы с увеличением основания кода m вероятность ошибочного приема стремилась к нулю. Используя более точные оценки, можно показать [17], что p_m стремится к нулю с ростом m, если

$$H'(A) < (P_c/N_0) \log e. \tag{4.64}$$

Но правая часть (4.64) есть не что иное, как пропускная способность (3.66) рассматриваемого канала при отсутствии ограничений на пропускаемую полосу частот. Таким образом, еще раз подтверждается теорема Шеннона: при $H'(A)<C$ можно, используя сигналы, передающие информацию о достаточно длинных отрезках сообщения, обеспечить сколь угодно малую вероятность ошибки.

Следует еще раз подчеркнуть, что при заданном ансамбле сигналов существует конечная минимальная вероятность ошибки, определяющая потенциальную помехоустойчивость. Если же ансамбль сигналов можно расширять, охватывая каждым элементарным сигналом все больший объем информации, то в условиях теоремы Шеннона можно получить сколь угодно низкую вероятность ошибки.

[1] Ортогональные сигналы в рассматриваемом канале не являются оптимальными. Однако доказано, что при больших m ортогональная система близка к оптимальной. Еще ближе к ней система с биортогональными сигналами, в которой к каждому сигналу имеется один противоположный, а остальные ему ортогональны.

Поскольку с ростом m вероятность ошибок уменьшается, может возникнуть вопрос, почему существующие системы передачи дискретных сообщений чаще всего двоичные, значительно реже $m=4$ и уж совсем редко используются значения $m>8$. Это объясняется, во-первых, тем, что с увеличением m быстро возрастает сложность демодулятора, значительно быстрее, чем уменьшается вероятность ошибки. Во-вторых, увеличение числа ортогональных сигналов требует расширения спектра, что не всегда возможно. Действительно, при ширине спектра F размерность пространства сигналов на интервале T приблизительно равна $2FT$ (см. § 2.4). Следовательно, любой ортогональный базис может содержать только $2FT$ элементов, так что $m\leqslant 2FT$. Поэтому с увеличением m должно пропорционально расти произведение FT. Так как T растет при этом пропорционально $\log m$, то ширина спектра F должна увеличиваться, по крайней мере, пропорционально $m/\log m$. Можно, конечно, отказаться от применения ортогональных сигналов и увеличивать m, не расширяя спектра. По теореме Шеннона и в этом случае существует возможность получить сколь угодно малую вероятность ошибок. Однако конструктивные методы непосредственного построения таких ансамблей аналоговых сигналов неизвестны. Именно вследствие этого приходится прибегать к построению сложных сигналов путем кодирования на дискретном уровне, которое будет рассмотрено в § 5.2.

4.6. ОПТИМАЛЬНЫЙ ПРИЕМ ПРИ СИГНАЛАХ С НЕОПРЕДЕЛЕННОЙ ФАЗОЙ (НЕКОГЕРЕНТНЫЙ ПРИЕМ)

Как было показано в § 3.4, многие каналы можно описать моделью (3.30) с флуктуирующей фазой. Естественно, если фаза (или какой-либо другой параметр) принимаемого сигнала флуктуирует настолько медленно, что путем измерения (оценки) ее можно достаточно надежно предсказать, оптимальный прием в основном реализуется так же, как при точно известном сигнале (с добавлением блоков оценки). Такая ситуация характерна для многих каналов проводной и, реже, радиосвязи. Однако нередко фаза флуктуирует довольно быстро, и точную ее оценку получить не удается. Кроме того, оценка фазы требует иногда применения сложных устройств. Поэтому даже в тех случаях, когда принципиально можно оценить начальную фазу приходящего сигнала, порой от этого отказываются и используют алгоритм, построенный в предположении, что начальная фаза приходящего сигнала неизвестна и может с равной вероятностью принимать любое значение на интервале 0; 2π. Такой метод приема называется *некогерентным*.

Для вывода оптимального алгоритма некогерентного приема будем исходить из отношения правдоподобия Λ_i для сигнала $s_i(t)$ (относительно нулевой гипотезы), которое при точно известной начальной фазе выражается формулой (4.25). Используя представление для сигнала (3.30), где k — известный коэффициент передачи канала, а $\theta_к$ — случайный сдвиг в канале, формулу (4.25) можно записать так:

$$\ln \Lambda_i = \frac{2k}{N_0}\int_0^T z(t)\,[u_i(t)\cos\theta_к + \tilde{u}_i(t)\sin\theta_к]\,dt - \frac{k^2}{N_0}\int_0^T [u_i(t)\cos\theta_к + \tilde{u}_i(t)\sin\theta_к]^2\,dt. \qquad (4.65)$$

Здесь $\ln \Lambda_i$ является случайной величиной, принимающей различные значения при различных $\theta_к$. Правило максимума правдоподобия в такой ситуации заключается в выборе такого решения, для которого математическое ожидание Λ_i будет наибольшим. Такой алгоритм кратко записывается так:

$$\max_i \bar{\Lambda}_i = \max_i \int_0^{2\pi} w(\theta_к)\, \Lambda_i(\theta_к)\, d\theta_к = \max_i \frac{1}{2\pi} \int_0^{2\pi} \Lambda_i(\theta_к)\, d\theta_к, \tag{4.66}$$

где $w(\theta_к) = 1 : 2\pi$ при $0 \leqslant \theta \leqslant 2\pi$ — плотность распределения вероятности $\theta_к$.

При нахождении Λ_i заметим, что второй интеграл в правой части (4.65) от $\theta_к$ не зависит и равен энергии $E_{пi}$ сигнала $u_i(t)$ на входе канала (на передатчике). Это ясно из того, что подынтегральной функцией является квадрат сигнала $u_i(t)$, сдвинутого по фазе на $\theta_к$, что, как известно, не влияет на его энергию. Таким образом, учитывая, что $k^2 E_{пi}/N_0 = E_i/N_0 = h^2_i$, получим

$$\bar{\Lambda}_i = \exp(-h_i^2) \frac{1}{2\pi} \int_0^{2\pi} \exp\left\{ \frac{2k}{N_0} \left[\cos\theta_к \int_0^T z(t)\, u_i(t)\, dt + \right.\right.$$

$$\left.\left. + \sin\theta_к \int_0^T z(t)\, \tilde{u}_i(t)\, dt \right]\right\} d\theta_к = \exp(-h_i^2) \frac{1}{2\pi} \int_0^{2\pi} \exp\left\{ \frac{2}{N_0} [y_i \cos\theta_к + \right.$$

$$\left. + \tilde{y}_i \sin\theta_к] \right\} d\theta_к, \tag{4.67}$$

где

$$y_i = k \int_0^T z(t)\, u_i(t)\, dt; \quad \tilde{y}_i = k \int_0^T z(t)\, \tilde{u}_i(t)\, dt. \tag{4.68}$$

Обозначив

$$V_i = \sqrt{y_i^2 + \tilde{y}_i^2} \quad \text{и} \quad \theta_i = \operatorname{arc\,tg} \tilde{y}_i / y_i, \tag{4.69}$$

можно записать

$$\bar{\Lambda}_i = \frac{1}{2\pi} \int_1^{2\pi} \exp\left\{ \frac{2}{N_0} V_i \cos(\theta_к - \theta_i) \right\} d\theta_к \exp(-h_i^2) = I_0\left(\frac{2V_i}{N_0}\right) \exp(-h_i^2), \tag{4.70}$$

где

$$I_0(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{x\cos(\varphi - \theta)}\, d\varphi \tag{4.71}$$

— модифицированная функция Бесселя (2.84).

Вместо того чтобы сравнивать отношения правдоподобия $\bar{\Lambda}_i$, можно сравнить их логарифмы, что приводит к следующему алгоритму *оптимального некогерентного приема:*

$$\max_i [\ln I_0(2V_i/N_0) - h_i^2]. \tag{4.72}$$

Величины y_i и $\tilde{y}_i$ можно получить в момент отсчета T на выходе активного фильтра с опорными сигналами, равными соответственно $u_i(t)$ и $\tilde{u}_i(t)$. С учетом сказанного понятно построение на основе активных фильтров схемы, называемой квадратурной и реализующей алгоритм (4.73) (рис. 4.16). Здесь Γ_i — генераторы опорных

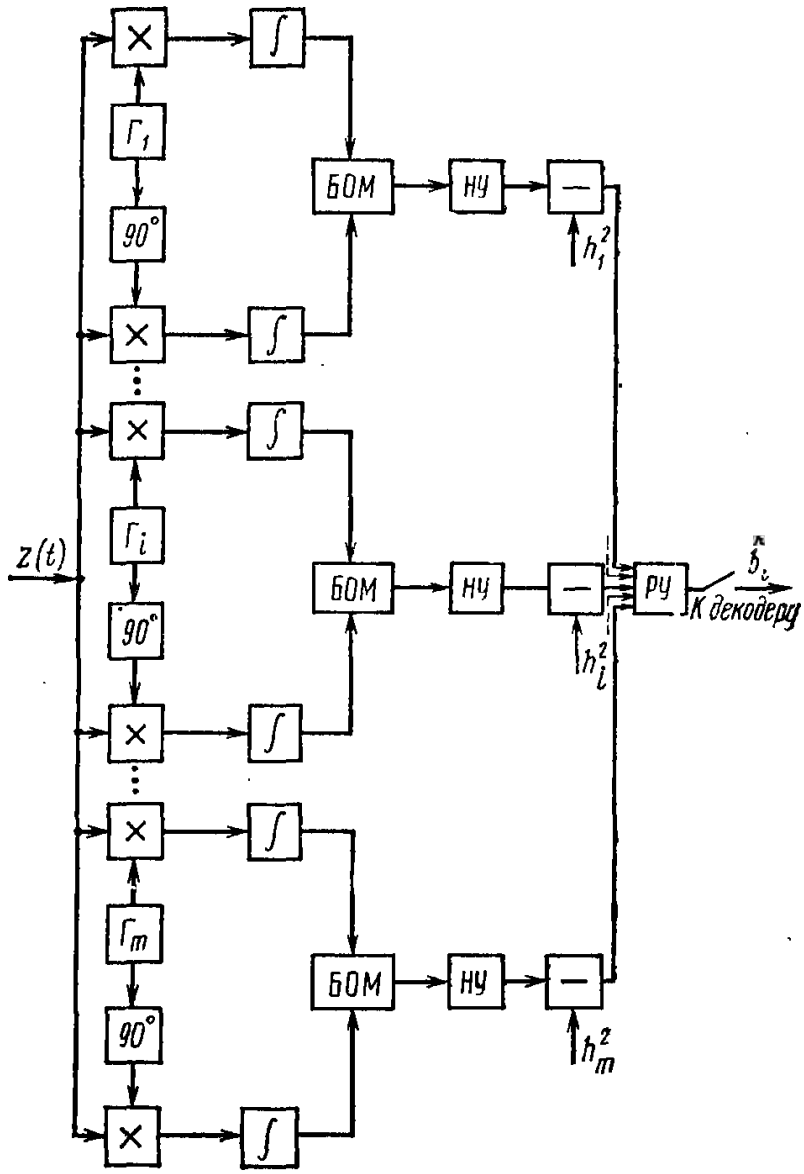


Рис. 4.16. Квадратурная схема реализации оптимального приема дискретных сообщений при неопределенной фазе сигнала

сигналов $u_i(t)$ с точностью до начальной фазы, 90° — фазовращатель всех сигнальных компонент на 90° (преобразователь Гильберта); БОМ — блок определения модуля вектора $V_i = \sqrt{y^2_i + \tilde{y}^2_i}$ по ортогональным компонентам; НУ — нелинейные безынерционные устройства с характеристикой

$$u_{вых} = \ln I_0 \left(\frac{2}{N_0} u_{вх} \right). \qquad (4.73)$$

Подчеркнем, что величины V_i не зависят от начальной фазы сигналов $u_i(t)$ и, как видно из (4.42), пропорциональны огибающей (в моменты отсчета, кратные T) на выходе фильтра, согласованного с сигналом $u_i(t)$. Таким образом, алгоритм (4.73) можно реализовать и на базе согласованных фильтров, как показано на рис. 4.17. Идеальный детектор Д выделяет огибающую напряжения на выходе согласованного фильтра.

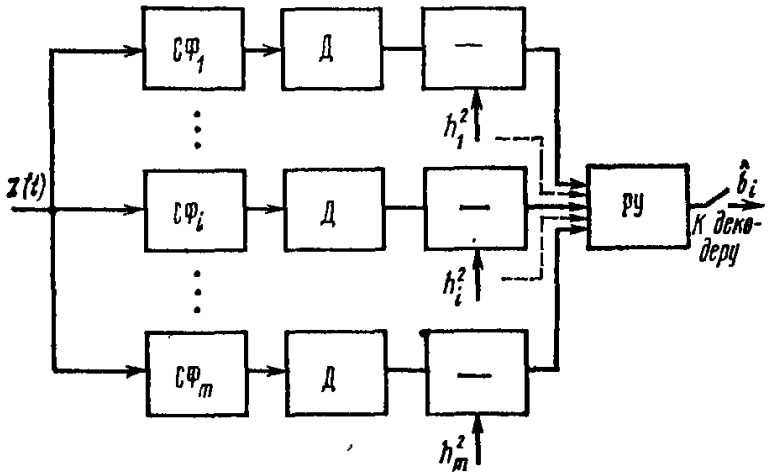


Рис. 4.17. Схема реализации оптимального приема дискретных сообщений на базе согласованных фильтров при неопределенной фазе сигнала

Для двоичной системы с пассивной паузой, полагая, что символ 0 передается сигналом $u_2(t)=0$, алгоритм (4.72) можно записать в виде

$$V_1 > \lambda, \tag{4.74}$$

где пороговый уровень

$$\lambda = \frac{N_0}{2} f(h^2), \tag{4.75}$$

а функция $x=f(y)$ обратна функции $y=\ln I_0(x)$. При выполнении неравенства (4.74) (превышении V_1 над порогом) регистрируется символ 1, в противном случае — символ 0.

Алгоритм (4.72) и соответственно его реализация существенно упрощаются для систем с активной паузой ($E_i=E$). Для них с учетом монотонного характера функции $\ln I_0(x)$ алгоритм оптимального некогерентного приема можно записать так:

$$\max_i V_i. \tag{4.76}$$

При его реализации в схемах рис. 4.16 и 4.17 не нужны ни блоки НУ, ни блоки вычитания. Более того, алгоритм при этом условии инвариантен относительно коэффициента передачи k и спектральной плотности шума N_0, поскольку V_i не зависит от N_0, а с изменением k все значения V_i изменяются пропорционально, что не влияет на (4.76). Именно это является основным преимуществом систем с активной паузой, определившим их широкое применение.

При выводе правила решения (4.72) предполагалось, что слу-

чайная начальная фаза распределена равномерно на интервале $(0; 2\pi)$. Однако в некоторых случаях распределение начальной фазы неравномерно, а еще чаще это распределение при построении системы связи неизвестно. В этих условиях возможны два подхода: а) построение адаптивной квазикогерентной системы, в которой путем анализа принимаемого сигнала определяется приближенная оценка фазы, используемая вместо недостающих априорных сведений, и б) принятие решения в предположении, что начальная фаза представляет собой некоторый неизвестный параметр, который так же, как и передаваемый кодовый символ, можно оценить по максимуму правдоподобия. Второй подход называют приемом по правилу *обобщенного максимума правдоподобия*.

Сущность этого правила заключается в следующем. Отношение правдоподобия для сигнала $u_i(t)$ при известном сдвиге фазы $\theta_к$ в канале записано в формуле (4.65). Найдем для данного i то значение $\theta_к$, которое обеспечивает максимум отношения правдоподобия $\max\limits_{\theta_к} \Lambda_i$ (или его логарифма), а затем сравним полученные значения для всех i и выберем из них наибольшее. Таким образом, приходим к алгоритму

$$\max_i \max_{\theta_к} \ln \Lambda_i. \qquad (4.77)$$

Для отыскания максимума (4.65) по $\theta_к$ учтем, что, как уже говорилось, второй интеграл от $\theta_к$ не зависит. Максимум же первого интеграла найдем обычным способом, продифференцировав его по параметру $\theta_к$ и приравняв производную нулю. Это приводит к уравнению $y_i \sin\theta_к + \tilde{y}_i \cos\theta_к = 0$, где y_i и $\tilde{y}_i$ определены формулами (4.68). Решая это уравнение, получим *максимально правдоподобное значение* $\hat{\theta}_к = \operatorname{arctg}(-\tilde{y}_i/y_i)$, откуда $\sin\hat{\theta}_к = -\tilde{y}_i/\sqrt{y^2_i+\tilde{y}^2_i}$; $\cos\hat{\theta}_к = y_i/\sqrt{y^2_i+\tilde{y}^2_i}$. Подставив эти значения (конечно, различные для разных гипотез) в (4.65), после простых преобразований получим следующий алгоритм решения по обобщенному максимуму правдоподобия:

$$\max_i \max_{\theta_к} \ln \Lambda_i = \max_i [V_i - E_i/2]. \qquad (4.78)$$

Для систем с активной паузой ($E_i = E = \text{const}$) это правило совпадает с (4.76). В этом случае алгоритм, полученный при неизвестной фазе, оцениваемой по максимуму правдоподобия, совпадает с алгоритмом, полученным в предположении, что фаза распределена равномерно.

Заметим попутно, что одной из актуальных проблем теории связи является отыскание алгоритмов решения для демодулятора, применимых при недостаточной априорной информации о канале и об источнике сообщения, например, об априорных вероятностях различных сигналов, о распределениях вероятностей амплитуд и фаз, о некоторых параметрах, входящих в описание сигнала и т. д. В этом направлении сделано уже очень много [12]. Конечно, чем больше объем априорной информации, тем достовернее можно принимать сообщение, например, применяя когерентный прием. Однако если сама априорная информация ненадежна, то, применяя алгоритм, учитывающий эту ненадежную информацию, можно получить результат хуже, чем при использовании алгоритма, построенного в предположении отсутствия данной априорной информации.

Исследование вероятности ошибок в канале с неопределенной фазой и аддитивным гауссовским шумом при поэлементном приеме показало, что минимальную вероятность ошибки обеспечивает

система с активной паузой, у которой сигналы удовлетворяют условиям ортогональности в усиленном смысле (см. § 2.4):

$$\int_0^T u_i(t)\,u_j(t)\,dt=0,\ i\neq j;\ \int_0^T u_i(t)\,\widetilde{u}_j(t)\,dt=0 \quad \text{при } i,\ j=1,\ 2,\ \ldots,\ m. \tag{4.79}$$

$$\int_0^T u_i^2(t)\,dt=E=\text{const}\,(i).$$

Можно построить много различных ансамблей сигналов, ортогональных в усиленном смысле. К ним относятся, например, сигналы ЧМ:

$$u_i(t)=a\cos\left[\frac{2\pi(l+i)}{T}t+\varphi\right],$$

где $i=1, 2, 3, \ldots, m$; l — любое целое число.

Другой пример представляют сигналы с временно́й манипуляцией (ВМ):

$$u_1(t)=\begin{cases}a(t);\ 0\leqslant t<T/2\\ 0;\ T/2\leqslant t<T;\end{cases}\qquad u_2(t)=\begin{cases}0;\ 0\leqslant t<T/2\\ a\left(t-\frac{T}{2}\right);\ \frac{T}{2}\leqslant t<T,\end{cases}$$

где $a(t)$ — любая функция из $L_2(T/2)$, в частности, она может быть отрезком гармонического сигнала.

Еще один пример:

$$u_1(t)=a\cos(2\pi l_1 t/T)+a\cos(2\pi l_2 t/T);\ u_2(t)=a\cos(2\pi l_1 t/T)-$$
$$-a\cos(2\pi l_2 t/T),$$

где l_1 и l_2 — любые целые числа.

Определим вероятность ошибки при приеме по алгоритму (4.76) двоичных сигналов, удовлетворяющих условиям (4.79). Если передается символ b_1, то с учетом (4.17), (4.68) и (4.69) имеем:

$$V_1=\sqrt{(\xi_1+E\cos\theta_к)^2+(\widetilde{\xi}_1+E\sin\theta_к)^2}; \tag{4.80}$$

$$V_2=\sqrt{\xi_2^2+\widetilde{\xi}_2^2}, \tag{4.81}$$

где $\xi_j=k\int_0^T N(t)\,u_j(t)\,dt;\ \widetilde{\xi}_j=k\int_0^T N(t)\,\widetilde{u}_j(t)\,dt,\ j=1,\ 2.$ (4.82)

Рассуждая так же, как при выводе формул (4.48) и (4.49), легко убедиться, что величины ξ_j, $\widetilde{\xi}_j$ распределены нормально с нулевым средним значением и дисперсией $0{,}5\,N_0E$.

Легко также показать, что коэффициенты корреляции $\overline{\xi_1\xi_2}$ и $\overline{\xi_1\widetilde{\xi}_2}$ при системе сигналов, ортогональной в усиленном смысле, равны нулю. Действительно, с учетом (4.82)

$$\overline{\xi_1\xi_2} = k^2 \int_0^T N(t)\,u_1(t)\,dt \int_0^T N(t)\,u_2(t)\,dt = k^2 \int_0^T \int_0^T \overline{N(t_1)\,N(t_2)}\,u_1(t_1) \times$$

$$\times\, u_2(t_2)\,dt_1 dt_2 = k^2 \int_0^T \int_0^T (N_0/2)\,\delta(t_2 - t_1)\,u_1(t_1)\,u_2(t_2)\,dt_1 dt_2 =$$

$$= (k^2 N_0/2) \int_0^T u_1(t)\,u_2(t)\,dt = 0. \tag{4.83}$$

Аналогично, $\overline{\xi_1\widetilde{\xi}_2} = \dfrac{k^2 N_0}{2} \int_0^T u_1(t)\,\widetilde{u}_2(t)\,dt = 0.$ (4.84)

Некоррелированность гауссовских величин обеспечивает и их независимость. Как следствие, случайные величины V_1 и V_2 независимы, причем V_2 в соответствии с (2.76) имеет распределения Рэлея

$$w(V_2) = \frac{2V_2}{N_0 E} \exp\left(-V_2^2/N_0 E\right),\ V_2 \geqslant 0, \tag{4.85}$$

V_1 в соответствии с (2.86) имеет обобщенное распределение Рэлея

$$w(V_1) = \frac{2V_1}{N_0 E} \exp\left(-\frac{V_1^2 + E^2}{N_0 E}\right) I_0\left(\frac{2V_1}{N_0}\right),\ V_1 \geqslant 0. \tag{4.86}$$

Вероятность приема символа b_2 при передаче символа b_1 определится вероятностью выполнения неравенства $V_2 > V_1$:

$$P(b_2 \mid b_1) = \int_0^\infty w(V_1) \int_{V_1}^\infty w(V_2)\,dV_2 dV_1 = \int_0^\infty \frac{2V_1}{N_0 E} \exp\left(-\frac{V_1^2 + E^2}{N_0 E}\right) \times$$

$$\times\, I_0\left(\frac{2V_1}{N_0}\right) \int_{V_1}^\infty \frac{2V_2}{N_0 E} \exp\left(-\frac{V_2^2}{N_0 E}\right) dV_1 dV_2 = \int_0^\infty \frac{2V_1}{N_0 E} \exp\left(-\frac{V_1^2 + E^2}{N_0 E} - \right.$$

$$\left. - \frac{V_1^2}{N_0 E}\right) I_0\left(\frac{2V_1}{N_0}\right) dV_1 = 0{,}5 \exp\left(-\frac{E}{2N_0}\right) \int_0^\infty \frac{4V_1}{N_0 E} \exp\left(-\frac{4V_1^2 + E^2}{2N_0 E}\right) \times$$

$$\times\, I_0\left(\frac{2V_1}{N_0}\right) dV_1. \tag{4.87}$$

В полученной формуле интеграл, на первый взгляд, очень сложен, но его можно легко вычислить с помощью теории вероятностей. Если заменить переменную, положив $a = 2V_1$, то окажется, что это интеграл от плотности вероятностей (2.86) случайной величины $A \geqslant 0$, имеющей обобщенное распределение Рэлея, с пара-

метрами $u=E$ и $\sigma^2=N_0E$. Так как интеграл берется по всей области определения A, то он равен 1. Окончательно

$$P(b_2 \mid b_1) = 0{,}5\exp(-0{,}5h^2), \qquad (4.88)$$

где, как и раньше, $h^2=E/N_0$ — отношение энергии элемента принимаемого сигнала к спектральной плотности мощности шума.

Из соображений симметрии ясно, что такова же будет вероятность приема символа b_1 при передаче b_2. Поэтому вероятность ошибки не зависит от передаваемого символа. Она одинакова для всех двоичных систем, ортогональных в усиленном смысле (при одинаковых энергиях сигнала), и определяется формулой (4.88). В частности, эта формула справедлива для системы ЧМ, для системы с временно́й манипуляцией и любых других систем, для которых выполняется (4.79).

На рис. 4.18 показана зависимость $p(h^2)$ согласно (4.88) (кривая *2*). Там же для сравнения приведена кривая, характеризующая потеициальную помехоустойчивость той же системы при коге-

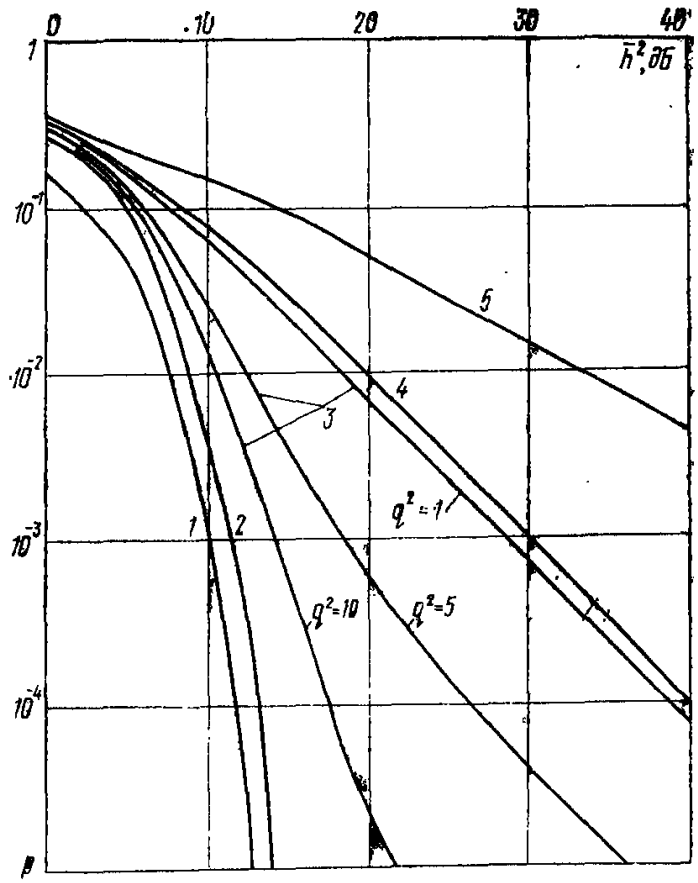


Рис. 4.18. Зависимости вероятности ошибки в двоичной системе с активной паузой, ортогональной в усиленном смысле (например, ЧМ), от h^2 при оптимальном приеме и различных параметрах канала:

1 — канал с постоянными параметрами (когерентный прием); *2* — канал с неопределенной фазой без замираний (некогерентный прием); *3* — обобщенные рэлеевские замирания: *4* — рэлеевские замирания; *5* — односторонне-нормальные замирания. Аддитивный шум во всех случаях белый

рентном приеме, определяемая формулой (4.57) (кривая *1*)[1]. Величина h для удобства выражена в децибелах, а вероятности ошибок отложены в логарифмическом масштабе.

Сравнение кривых показывает, что для рассматриваемой системы связи (с активной паузой, ортогональной в усиленном смысле)

[1] Об остальных кривых на этом рисунке будет сказано ниже.

априорное знание фазы и осуществление когерентного приема дают лишь очень небольшой энергетический выигрыш по сравнению со случаем некогерентного приема. Этот выигрыш тем меньше, чем ниже допустимая вероятность ошибки.

При основании кода $m>2$ для систем с активной паузой ($E_i=E$), ортогональных в усиленном смысле, при оптимальном некогерентном приеме вероятность ошибок выражается довольно сложной формулой [17]. При небольших вероятностях достаточно плотную оценку можно получить из неравенства Буля:

$$p_m \leqslant (m-1)\, p_2 = 0{,}5\,(m-1)\exp(-0{,}5\,h^2). \qquad (4.89)$$

Систему ФМ, так же как и другие системы с противоположными сигналами, отличающимися сдвигом фаз на π, при некогерентном приеме применять нельзя, так как при неизвестной начальной фазе такие сигналы неразличимы. Однако разности фаз между двумя элементами сигнала различаются, если сдвиг фазы в канале изменяется достаточно медленно. Поэтому вполне возможен некогерентный прием при ОФМ.

Поскольку при ОФМ информационный параметр сигнала определяется двумя соседними элементами [(n—1)-м на интервале $-T \div 0$ и n-м на интервале 0—T], то оптимальный алгоритм (4.76) можно записать в виде

$$\max_i \left\{ \sqrt{\left[\int_{-T}^{T} z(t)\, u_i(t)\, dt\right]^2 + \left[\int_{-T}^{T} z(t)\, \widetilde{u}_i(t)\, dt\right]^2} \right\}, \quad i=1,\ 2. \qquad (4.90)$$

Приходящий сигнал $s(t)$ на двух тактовых интервалах при ОФМ можно представить в зависимости от символа, передаваемого n-м элементом, так:

$$\left.\begin{aligned} s_1(t) &= a\cos(\omega_0 t+\psi),\ 0 \leqslant t < 2T \ \text{(при передаче символа 1)};\\ s_2(t) &= \left\{\begin{array}{l} a\cos(\omega_0 t+\psi),\ 0 \leqslant t < T \\ -a\cos(\omega_0 t+\psi),\ T \leqslant t < 2T \end{array}\right\} \text{(при передаче символа 0)}, \end{aligned}\right\} \qquad (4.91)$$

где ψ — случайная начальная фаза, неизвестная при приеме, зависящая, в частности, от символа, передававшегося (n—2)-м элементом.

Нетрудно видеть, что (4.91) представляет собой двоичную систему сигналов с активной паузой, ортогональную в усиленном смысле на интервале длительностью $2T$, а не T. Поэтому вероятность ошибки при приеме сигналов ОФМ по алгоритму (4.90) определяется на основании (4.88), но с учетом того, что энергия сигнала на интервале — T÷T равна $2E$:

$$p = 0{,}5\exp(-h^2), \qquad (4.92)$$

где параметр h^2 — отношение энергии сигнала на интервале длительностью T к спектральной плотности шума. Как и следовало ожидать, вероятность ошибки (4.92) несколько больше, чем вычисленная для когерентного приема двоичной ОФМ (4.59), однако различие между ними совершенно ничтожно.

Для схемной реализации алгоритм (4.90) можно упростить. Для этого подставим систему сигналов (4.91) в (4.90) и после сокращения одинаковых слагаемых приведем алгоритм приема к виду

$$X_a X_b + Y_a Y_b > 0, \tag{4.93}$$

где

$$\left.\begin{aligned} X_a &= \int_{-T}^{0} z(t)\cos(\omega_0 t+\psi)\,dt;\quad X_b = \int_{0}^{T} z(t)\cos(\omega_0 t+\psi)\,dt;\\ Y_a &= \int_{-T}^{0} z(t)\sin(\omega_0 t+\psi)\,dt;\quad Y_b = \int_{0}^{T} z(t)\sin(\omega_0 t+\psi)\,dt. \end{aligned}\right\} \tag{4.94}$$

Полагая фазу ψ хотя и случайной, но постоянной на интервале — $T \div T$, можно легко показать, что левая часть (4.93) инвариантна к значению этой фазы.

На рис. 4.19 показана корреляционная схема, реализующая алгоритм приема (4.93) на основе активных фильтров. Величины

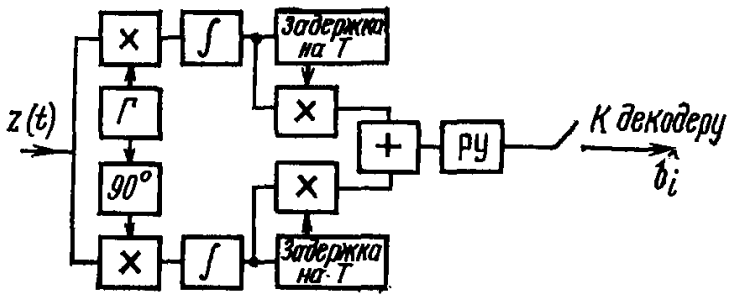


Рис. 4.19. Схема оптимального некогерентного приема сигналов ОФМ на базе активных фильтров

X_a, X_b, Y_a, Y_b получаются путем интегрирования произведения элемента принимаемого колебания на опорные сигналы $\cos(\omega_0 t+\psi)$ и $\sin(\omega_0 t+\psi)$ на интервале длительностью T.

В моменты времени, кратные T, величины X_b и Y_b снимаются непосредственно с интеграторов, а X_a и Y_a — с выхода цепи задержки на время T. На рис. 4.19 не показаны цепи, осуществляющие сброс интегратора к концу интервала интегрирования и ввод накопленного на нем результата в перемножитель и цепь задержки на время T.

Структурная схема, реализующая алгоритм (4.93) на основе коммутируемых фильтров и используемая в различных вариантах систем с ОФМ, показана на рис. 4.20.

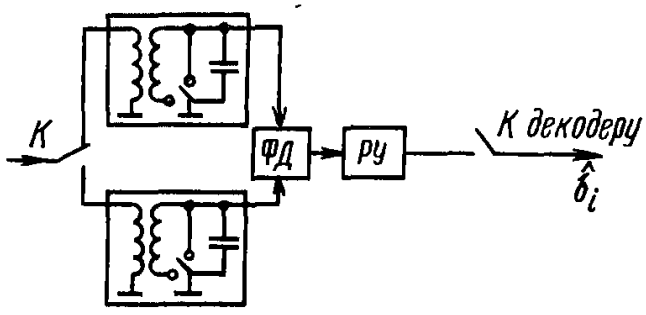


Рис. 4.20. Схема оптимального некогерентного приема сигналов ОФМ на основе коммутируемых фильтров

В момент $t=-T$ прнходящее колебанне $z(t)$ подается на первый фнльтр (резонансный контур с высокой добротностью), который до этого приведен к нулевым начальным условиям путем гашення колебанй. В момент $t=0$ входное колебание переключается на второй фильтр, тоже приведенный к нулевым начальным условиям. В первом фильтре продолжаются собственные колебания до момента отсчета T. Колебания с двух фильтров поступают на фазовый детектор ФД (выполняющий функцню перемножения входных снгналов н интегрирования результата). В момент $t=2T$ входной сигнал опять подключается к первому фнльтру, а колебания во втором фильтре продолжаются до момента $3T$ и т. д. Можно показать, что знак напряженння на выходе фазового детектора в моменты, кратные T, позволяет прииять нужное решение. Существуют и другие схемы некогерентного приема сигналов с ОФМ.

При некогерентной обработке высокочастотных сигналов (обработке по огибающей) снижаются требования к точности установки границ посылок элементарных канальных сигналов длительностью T. Все же для реализации оптимальной схемы средняя частота заполнения сигналов должна быть известна с высокой точностью. Во всяком случае, если отклонение средней частоты δF от номинального значения f_0 существенно меньше $1/T$:

$$|\delta F| \ll 1/T, \tag{4.95}$$

то «набег фазы» средней частоты на интервале T (т. е. величина $2\pi|\delta F \cdot T|$) существенно меньше 2π и можно пользоваться полученными выше алгоритмами оптимального приема и формулами для минимальной вероятности ошибок.

Остановимся кратко на некоторых схемах неоптнмального приема при неопределенной фазе снгнала, широко используемых в современной аппаратуре связн. При приеме снгналов двоичной АМ распространена схема рис. 4.21. Здесь амплитудный детектор Д н фильтр ннжних частот ФНЧ выделяют огибающую $r(t)$ приннмаемого колебання, прошедшего входной избирательный блок (полосовой фильтр ПФ) с эффективной полосой пропускания $F_э$, достаточной для получения всех наиболее существенных компонент сигнала. Огибающая $r(t)$ с выхода ФНЧ в определенные моменты времени (например, в середине посылкн) сравнивается в РУ с некоторым пороговым уровнем λ. При выполненнн неравенства

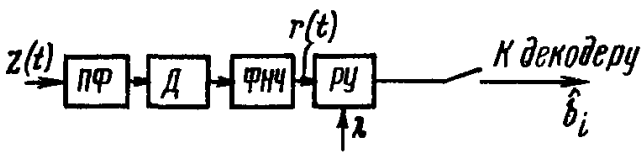


Рис. 4.21. Схема неоптимального приема снгналов АМ методом сравнения огибающей с пороговым уровнем

$$r > \lambda \tag{4.96}$$

регнстрируется символ 1, в противном случае — 0. Сравнивая (4.96) с алгоритмом (4.74), можно видеть, что схема рис. 4.21 отличается от оптимальной некогерентной схемы рис. 4.17 использованием полосового фнльтра и последетекторного фильтра нижних частот, вместо одного согласованного фильтра до детектора.

При приеме сигналов двоичной ЧМ распространена схема рис. 4.22, где $ПФ_1$ н $ПФ_0$ — разделительные полосовые фильтры, пропускающие без существенных искаженнй соответственно сигналы $s_1(t)$ и $s_0(t)$; Д — амплитудный детектор. Разностный сигнал двух детекторов подвергается фильтрацин в ФНЧ, а результат для выбора решения сравнивается с нулевым порогом.

Анализ такой схемы приводит к следующим результатам. Вероятность ошибок в схеме рис. 4.22 больше, чем при оптимальном некогерентном приеме, причем ее возрастание обусловливается двумя основными факторами:

а) уменьшением отношения мощности сигнала к мощности шума по сравненню с согласованным фильтром;

б) межсимвольными помехами, создаваемыми переходными процессами в фильтрах (остаточными собственными колебаниями, возникшими в результате воздействия предыдущих элементов сигнала).

Как указывалось в § 4.4 (см. с. 138), первый из этих факторов вызывает наименьшую потерю помехоустойчивости, если полоса пропускания полосового фильтра $\Delta f = 1.37/T$. Однако при такой полосе пропускания весьма существенные

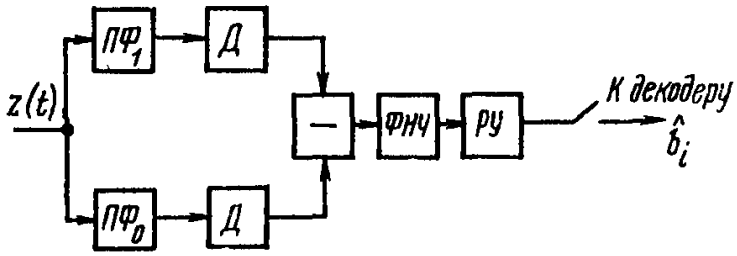


Рис. 4.22. Схема неоптимального некогерентного приема сигналов ЧМ с разделительными полосовыми фильтрами

погрешности вносятся за счет второго фактора — межсимвольной интерференции. Поэтому наименьшая вероятность ошибок в схеме с полосовыми разделительными фильтрами при отсутствии ФНЧ достигается при более широкой полосе пропускания, примерно при $\Delta f \approx 3/T$. Можно показать, что для получения одинаковой вероятности ошибок в схеме с полосовыми разделительными фильтрами требуется в ΔfT раз (в данном примере в 3 раза) бо́льшая мощность сигнала, чем в схеме оптимального некогерентного приема, что и определяет энергетический проигрыш при замене согласованных фильтров полосовыми.

Схемы приемников с неоптимальной фильтрацией до и после детектора широко используются на практике в тех случаях, когда частотная стабильность недостаточна, т. е. условие (4.95) $|\delta F| \leqslant 1/T$ не выполняется, и, следовательно, реализация оптимального приема с согласованной фильтрацией фактически невозможна. Это имеет место, например, вследствие эффекта Допплера при связи с движущимися объектами или при использовании движущегося спутника для ретрансляции при больших нестабильностях частот автогенераторов и т. п.

Если полосы пропускания входных фильтров Δf в схемах рис. 4.21 и 4.22 удовлетворяют условию $\Delta f > 2|\delta F|$, то сигнал остается в полосе пропускания фильтра при всевозможных флуктуациях частот. При этом величина ΔfT может оказаться значительно больше 1, и не будь фильтрации сигнала после детектора, энергетический проигрыш по сравнению с оптимальным приемом при стабильной частоте сигнала был бы весьма существен. Однако есть возможность значительно уменьшить этот проигрыш, применив фильтрацию напряжения, снимаемого с выхода детектора. При $2|\delta F|T \gg 1$ полосовой фильтр почти не искажает огибающую входного сигнала, поэтому при отсутствии помех напряжение на выходе детектора в схеме рис. 4.21 представляет собой однополярные видеоимпульсы, а на выходе блока вычитания схемы рис. 4.22 — двухполярные. При небольшом уровне шума на входе детектора условия на его выходе приближенно такие же, как при приеме прямоугольных видеоимпульсов на фоне белого гауссовского шума. Поэтому естественно включить после детектора фильтр, согласованный с прямоугольным импульсом (рис. 4.10*а*), либо эквивалентный ему коммутируемый интегратор (рис. 4.4). На практике часто применяют и несогласованный последетекторный фильтр нижних частот, как это показано в схемах рис. 4.21, 4.22.

Теория таких схем довольно сложна. Она показывает, что при использовании полосовых разделительных фильтров и согласованного ФНЧ прием сигналов ЧМ осуществляется с вероятностью ошибки, ненамного большей, чем при оптимальном некогерентном приеме. Энергетический проигрыш оценивается здесь величиной 1—2 дБ. При несогласованном, но достаточно узкополосном ФНЧ

энергетический проигрыш больше, но обычно не превышает 4—5 дБ. Роль ФНЧ определяется тем, что его полоса пропускання может выбираться независимо от стабильностн несущей частоты. Поэтому последетекторная фнльтрацня в некоторой степени окупает недостаточную (вследствие нестабильностн частоты) фнльтрацию до детектора.

4.7. ПРИЕМ ДИСКРЕТНЫХ СООБЩЕНИЙ В КАНАЛАХ С ЗАМИРАНИЯМИ. РАЗНЕСЕННЫЙ ПРИЕМ

В большей части радиоканалов, а также в некоторых других каналах флуктуирует не только начальная фаза, но и амплитуда ожидаемых сигналов $s_i(t)$ (коэффициент K). При относительно быстрых (по сравнению с длительностью посылки T) замираниях сигнала нельзя сколь-нибудь определенно судить по результатам приема предыдущих элементов о значениях амплитуд и фаз последующих элементов.

Пусть канал описывается моделью (3.32), т. е. является одно-лучевым гауссовским с общими замираниями. Алгоритм оптимального приема в этих условиях нетрудно получить, определив математическое ожидание от (4.70) по k:

$$\overline{\Lambda}_i = \int_0^\infty \overline{\Lambda}_i(k)\, w(k)\, dk \tag{4.97}$$

и сравнив между собой отношения правдоподобия $\overline{\Lambda}_i$ с различными индексами i. Однако для систем с активной паузой результат легко указать и без дополнительных выкладок — он определяется соотношением (4.76). Это очевидно, так как (4.76), являясь алгоритмом приема при неопределенной фазе, не зависит от амплитуды (коэффициента K), следовательно, этот алгоритм остается оптимальным при любом законе распределения амплитуд. При этом, однако, помехоустойчивость приема существенно зависит от распределения K.

Определим, например, вероятность ошибки для двоичной системы с активной паузой, ортогональной в усиленном смысле, при условии, что замирания в канале медленные.

Если условную вероятность ошибки при некотором фиксированном значении k обозначить p_k, то безусловная вероятность ошибки при медленных замираниях

$$p = \int_0^\infty p_k w(k)\, dk. \tag{4.98}$$

В нашем случае условная вероятность ошибки определяется формулой (4.88), в которой величина $h = \sqrt{E/N_0} = \sqrt{k^2 E_п/N_0}$ пропорциональна k. Здесь $E_п$ — энергия сигнала на входе канала (на передатчике). Очевидно, что между значениями h, k и математическими ожиданиями их квадратов существует зависимость

$$h^2 = k^2 \overline{h^2}/\overline{K^2}. \tag{4.99}$$

Пусть, например K имеет распределение Рэлея (3.25), которое можно представить в следующей форме:

$$w(k)=\frac{2k}{\overline{K^2}}\exp(-k^2/\overline{K^2}).$$

Подставив это в (4.98) и обозначив $k^2/\overline{K^2}=x^2$, с учетом (4.88) и (4.99) найдем вероятность ошибки для двоичных ортогональных в усиленном смысле сигналов при рэлеевских замираниях:

$$p=\int_0^\infty 0{,}5\exp(-0{,}5k^2\overline{h^2}/\overline{K^2})\frac{2k}{\overline{K^2}}\exp(-k^2/\overline{K^2})\,dk=$$

$$=\int_0^\infty (k/\overline{K^2})\exp[-k^2(1+0{,}5\overline{h^2})/\overline{K^2}]\,dk=$$

$$=\int_0^\infty x\exp[-x^2(1+0{,}5\overline{h^2})]\,dx=1/(\overline{h^2}+2). \qquad (4.100)$$

Эта зависимость представлена на рис. 4.18 (кривая *4*).

Аналогично определяется вероятность ошибок и при других законах замираний. Так, например, если распределение вероятностей K — обобщенное рэлеевское (3.26), то

$$p=\frac{1+q^2}{2+2q^2+\overline{h^2}}\exp\left(-\frac{\overline{h^2}\,q^2}{2+2q^2+\overline{h^2}}\right), \qquad (4.101)$$

где q^2 — отношение мощностей регулярной и флуктуирующей составляющих. На рис. 4.18 (кривые *3*) показана эта зависимость при $q^2=1$; 5 и 10. Легко проверить, что при $q^2=\infty$ (отсутствие замираний) формула (4.101) переходит в (4.88), а при $q^2=0$ (отсутствие регулярной составляющей) — в формулу (4.100).

Приведем еще результат для случая, когда случайная величина K распределена по одностороннему нормальному закону:

$$w(k)=\begin{cases}\sqrt{2/\pi\,\overline{K^2}}\exp(-k^2/(2\overline{K^2})) & \text{при } k\geqslant 0;\\ 0 & \text{при } k<0.\end{cases}$$

Это имеет место в очень плохих радиоканалах. При этом

$$p=1/\left(2\sqrt{1+\overline{h^2}}\right) \qquad (4.102)$$

(см. рис. 4.18, кривая *5*).

Для общего случая обобщенного гауссовского канала выражения вероятности ошибок можно найти в [8].

Заметим, что все полученные для двоичных систем выражения вероятностей ошибок — (4.51), (4.56)—(4.59), (4.88), (4.92), (4.100), (4.101), (4.102) — при h^2 (или $\overline{h^2}$), стремящемся к нулю, принимают значение 0,5. Это и следовало ожидать, так как при $p=0{,}5$ по двоичному каналу никакая информация не передается (см. рис. 3.9 и относящиеся к нему пояснения). При $h^2\to\infty$ вероятность ошибок стремится к нулю. Это значит, что во всех рассмотренных каналах можно получить сколь угодно малую вероятность

ошибки, увеличивая мощность сигнала. Однако степень этого увеличения различна для разных каналов. Сравнение кривых на рис. 4.18 показывает, что при замираниях сигнала помехоустойчивость систем связи значительно ниже, чем в канале без замираний, при той же средней мощности передатчика. Для поддержания заданного качества связи в этих условиях приходится иметь определенный запас мощности передатчика. Так, для того чтобы добиться вероятности ошибки $p=10^{-4}$ в канале с рэлеевскими замираниями, необходимо увеличить мощность, по сравнению с каналом без замираний, в 588 раз (на 27,7 дБ), а в односторонне нормальном канале — более чем в миллион раз (на 61,7 дБ).

С ростом регулярной составляющей помехоустойчивость связи монотонно возрастает. При $q^2>(10\div20)$ помехоустойчивость почти такая же, как в канале без замираний.

Вероятности ошибок в недвоичных ортогональных системах при замираниях можно оценивать с помощью неравенства (4.60). Впрочем, для многих практически важных случаев известны и точные формулы. Так, при рэлеевских замираниях для системы из m сигналов, ортогональных в усиленном смысле и имеющих одинаковую энергию,

$$p=1-\prod_{i=1}^{m-1}\frac{i}{i+(\overline{h^2}+1)^{-1}}. \qquad (4.103)$$

Поскольку интервал корреляции замираний обычно значительно больше длительности тактового интервала, ошибки в каиале с замираниями не являются независимыми. Если уровень приннмаемого сигнала на некотором тактовом интервале низок, то с большой вероятиостью он останется невысокнм и на нескольких последующих тактовых интервалах. Поэтому имеется тенденция к группированию ошибок.

Прн очень медленных замнраниях, когда иитервал корреляции соизмерим с длнтельностью сеаиса связи, определеиие вероятности ошибок по формуле (4.98) лишено смысла, так как средняя вероятиость ошнбок ие характеризует условий прнема отдельных сообщений. В этой снтуации, а также в тех случаях, когда уровень помех в канале подвержен медлеиным изменениям, удобнее характеризовать систему связи не средней вероятностью ошибок, а *надежностью*. Под надежностью поиимают вероятность того, что на протяженин некоторой условной длительности сеанса (обычно 5 или 10 мин) вероятность ошибки не превысит определениой заданной величины.

В более сложных непрерывных каналах со случайной структурой, описываемых моделью (3.33), задачи нахождения оптимального демодулятора, выбора оптимального ансамбля сигналов[1] и вычисления вероятностей ошибок представляют значительные трудности и в настоящее время еще окончательно не решены. Как уже отмечалось, в зависимости от структуры канала, используемой полосы частот и длительности элемента сигнала наблюдаются многолучевой характер распространения и селективные замирания.

[1] В отличие от ранее рассмотрениых каналов, здесь вероятность ошибок определяется не только средней энергией сигналов и их скалярными произведениями, но и «тонкой структурой» сигналов, например шириной спектра сигнала, автокорреляционными функциями и т. д.

Если в таком канале используется модем, рассчитанный без учета указанных особенностей, то селективные замирания и многолучевость (эхо-сигналы) приводят к межсимвольной интерференции (наложению растянувшихся элементов сигнала друг на друга) и к существенному увеличению вероятностей ошибок. Однако за последние два десятилетия были предложены некоторые новые методы построения модемов, учитывающие особенности такого канала. При их применении оказывается возможным использовать информацию, переносимую каждым лучом, и обеспечить в многолучевом канале более высокую верность, чем в однолучевом канале.

Известны различные подходы к задаче построения системы связи, позволяющей передавать дискретные сообщения с высокой вериостью и достаточной скоростью в каналах с многолучевым распространением. Один из них заключается в одновременной передаче нескольких узкополосных сигналов на разных частотах в общей достаточно широкой полосе частот, когда для каждого из частичных сигналов канал проявляется только общими замираниями. По этому методу строятся так называемые *параллельные* модемы. Требуемая достаточно высокая скорость передачи информации достигается увеличением основания кода ($m>2$) и передачей большого числа параллельных частичных сигналов.

При другом подходе (в *последовательных* модемах) применяются сигиалы с малым тактовым нитервалом, а межсимвольная интерференция, вызваиная многолучевым распространением, компенсируется путем нспользования информации о предыдущих элементах снгнала.

Для каналов с резко выраженной дискретной многолучевостью предложеиы и частично используются системы с *широкополосными шумоподобными сигналами,* имеющими базу $2FT \gg 1$. Поскольку нитервал корреляции $\tau_к$ нмеет величину порядка $1/F$, для них верио неравенство $\tau_к \ll T$. Это позволяет разделить в демодуляторе отдельиые лучн. Так, если демодулятор построен на согласованных фильтрах (иапример, рис. 4.8), то каждый из приходящнх лучей создаст на выходе соответствующего фильтра отклик с шириной пика порядка $\tau_к$, и если разность хода лучей больше $\tau_к$, то каждый из них можно выделить для дальнейшей обработки. Известны и другие схемы разделения лучей. Методы формирования широкополосных шумоподобных сигналов описаны иижe (см. § 8.4).

Каждый из перечисленных методов построения модема, а также и некоторые другие методы иаходят применение прн тех или нных характеристиках канала. Дальнейшие подробности о передаче информации в каналах с селективиыми и быстрыми замираниями можно найти в дополнительной литературе [17].

Вероятность ошибок при приеме дискретных сообщений можно существенно уменьшить с помощью *разнесенного приема*, сущность которого заключается в том, что демодулятор принимает решение о переданном символе ие по одному, а по двум или более сигналам, несущим одну и ту же информацию. Разнесенный прием является одним из основных способов повышения помехоустойчивости связи при замираниях сигнала.

В радиосвязи применяются различные способы разнесеиного приема: по времени (он сводится к повторению сигнала несколько раз); по частоте (сигнал дублируется по многим частотным каналам); прием сигнала на различные антенны, разнесенные в пространстве; поляризациоиное разнесение (прием на антенны, расположеиные в одном месте, но принимающне электромагнитные волны разной поляризации); разнесение по отдельным лучам в многолучевом канале. В последнем случае лучи разделяются либо остро иаправленными аитениами по углу прихода в горизонтальной или вертикальной плоскости (применяется, главным образом, иа УКВ), либо по времени прихода (временн запаздывания).

Как указывалось выше, разделенне лучей по времени прихода требует применения широкополосиых сигналов.

Из перечисленных методов в радиосвязи наибольшее распространенне получил прием на разнесенные в пространстве антенны. На втором месте стоит разнесенный прием по частоте, используемый в различных каналах. Остальные методы разнесения также находят применение, но значительно реже.

В каналах без замираний разнесенный прием повышает верность, если имеется возможность сложить пришедшие по n вствям сигналы *когерентно*, т. е. сводя имеющиеся между ними разности начальных фаз к пренебрежимо малой величине.

При когерентном сложении n одинаковых сигналов суммарный сигнал будет иметь в n раз бо́льшую «амплитуду», т. е. в n^2 раз бо́льшую мощность, чем отдельный сигнал. При этом помехи, которые обычно в различных ветвях независимы, складываются некогерентно, так что мощность суммарной помехи будет только в n раз больше мощности помехи в одной ветви. В результате отношение мощности сигнала к мощности помехи увеличится в n раз.

Можно показать, что если по n ветвям приннмаются сигналы с различными мощностямн, а помехи, присутствующие в них, имеют различную интенсивность, то наилучшие результаты получаются при когерентном сложении сигналов, умноженных на весовые коэффициенты, пропорциональные $h_r = \sqrt{E_r/N_{0r}}$, где E_r — энергия элемента приходящего сигнала, а N_{0r} — спектральная плотность шума в r-й ветви ($r=1,..,n$). При этом в суммарном сигнале отношение $h_э^2$ энергии сигнала к спектральной плотности шума равно

$$h_э^2 = \sum_{r=1}^{n} h_r^2. \qquad (4.104)$$

В каналах с замираниями имеется и другой, более эффективный механизм повышения верности при разнесенном приеме. Его можно грубо пояснить так: при одиночном приеме ошибки возникают, главным образом тогда, когда уровень сигнала упадет ниже некоторого порогового значения, а при разнесенном приеме — когда уровень сигнала окажется ниже порогового во всех ветвях. Если замирания в ветвях слабо коррелированны, то вероятность одновременного падения уровней сигнала во всех ветвях очень мала.

Существуют различные способы комбинирования («сложения») сигналов отдельных ветвей разнесения при приеме. Не останавливаясь на исследовании оптимальных способов сложения (разнесенного приема) в каналах с замираниями, рассмотрнм наиболее простой, достаточно эффективный и широко распространенный способ *автовыбора* ветви с наиболее сильным сигналом (рис. 4.23). В этой схеме постоянно измеряется коэффнциент передачи канала K_r (нли мощности принимаемого сигнала) по отдельным ветвям (приемникам), а к демодулятору подключается приемник с наиболее сильным сигналом.

Проанализируем помехоустойчивость схемы автовыбора для двоичной системы с активной паузой, ортогональной в усиленном смысле (например, ЧМ), при медленных рэлеевских независимых и идентичных ($\overline{K^2}_r = \overline{K^2}$) замираниях сигналов в отдельных ветвях.

Пусть $K_м$ — максимальное значение коэффициентов передачи K_r по всем ветвям ($r=1,\ldots,n$). Найдем функцию распределения вероятности случайной величины $K_м$, т. е. вероятности того, что оиа меньше некоторой величины k_0. Очевидно, что максимальное из значений реализаций k_r меньше k_0 тогда и только тогда, когда все значения k_r меньше k_0, т. е.

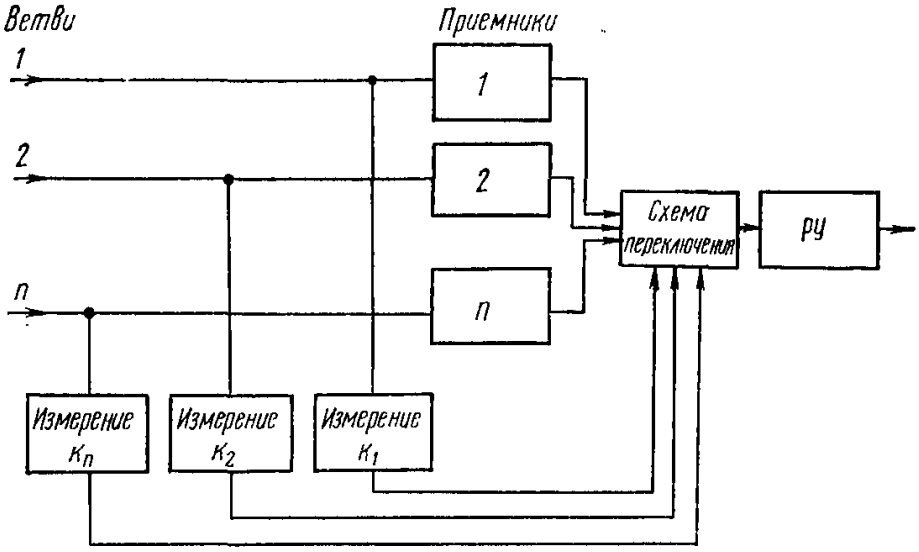


Рис. 4.23. Схема разнесенного приема при автовыборе ветви с наиболее сильным сигналом

$$F(k_0) = P\{K_м \leqslant k_0\} = P\{K_1 < k_0;\ K_2 < k_0;\ \ldots;\ K_n < k_0\} = \left[\int_0^{k_0} w(k)\,dk\right]^n =$$

$$= \left[\int_0^{k_0} \frac{2k}{\overline{K^2}} \exp\left(-\frac{k^2}{\overline{K^2}}\right) dk\right]^n = \left[1 - \exp\left(-\frac{k_0^2}{\overline{K^2}}\right)\right]^n. \qquad (4.105)$$

Третье равенство вытекает из предположения о независимости замираний в ветвях (независимости величин K_r). Здесь $w(k) = (2k/\overline{K^2})\exp(-k^2/\overline{K^2})$ — плотность вероятностей коэффициентов передачи при рэлеевских замираниях; $\overline{K^2}$ — средний квадрат коэффициента передачи, который мы считаем одинаковым во всех ветвях.

Для нахождения плотности распределения $K_м$ нужно $F(k_0)$ продифференцировать по k_0 и заменить k_0 на $k_м$:

$$w(k_м) = \frac{d}{d\,k_м}\left[1 - \exp\left(-\frac{k_м^2}{\overline{K^2}}\right)\right]^n =$$

$$= \frac{2n\,k_м}{\overline{K^2}} \exp\left(-\frac{k_м^2}{\overline{K^2}}\right)\left[1 - \exp\left(-\frac{k_м^2}{\overline{K^2}}\right)\right]^{n-1}. \qquad (4.106)$$

Для двоичной ортогональной системы в отсутствие замираний сигнала и при оптимальном некогерентном приеме вероятность ошибки при $k = k_м$ согласно (4.88) $p_{k_м} = 0{,}5\exp(-0{,}5k^2_м E_п/N_0)$, где N_0 — спектральная плотность шума; $k^2_м E_п = E$ — энергия элемента сигнала в выбранной ветви с $k = k_м$. Анализируемая схема автовыбора может рассматриваться как схема одиночного приема в эквивалентном канале, в котором при медленных замираниях сигнала коэффициент передачи $K = K_м$ меняется в соответствии со статистикой (4.106), т. е. средняя вероятность ошибки при n-кратном разнесении[1]

[1] Эта формула приближенна, так как из-за аддитивного шума в канале блоки измерения на рис. 4.23 не могут точно измерить мощность чистого сигнала (коэффициенты k_r). Однако при достаточно медленных замираниях, производя усреднение по помехе (выбирая достаточно большую постоянную времени), блоки измерения могут удовлетворительно справиться со своей задачей.

$$p_n = \int_0^\infty p_{h_м}\, w\,(k_м)\, dk_м = \frac{n}{\overline{K^2}} \int_0^\infty k_м \exp\left[-\frac{k_м^2}{\overline{K^2}}\left(1 + \right.\right.$$

$$\left.\left. + \frac{\overline{h^2}}{2}\right)\right]\left[1 - \exp\left(-\frac{k_м^2}{\overline{K^2}}\right)\right]^{n-1} dk_м,$$

$\overline{h^2} = \overline{K^2} E_п / N_0$ — отношение средней (по замираниям) энергии посылки сигнала в месте приема к спектральной плотности мощности шума.

Используя формулу бинома Ньютона и интегрируя, получим

$$p_n = \frac{n!}{2\prod_{r=1}^{n}(r + 0{,}5\overline{h^2})}. \qquad (4.107)$$

На рис. 4.24 изображены зависимости p_n от $\overline{h^2}$ при числе ветвей разнесения $n=2, 3, 4$. Там же для сравнения показана кривая для вероятности ошибок при неразнесенном (одиночном) приеме ($n=1$).

Из кривых видно, что эффективность разнесения велика при переходе от одиночного к сдвоенному приему и заметно менее выражена при дальнейшем

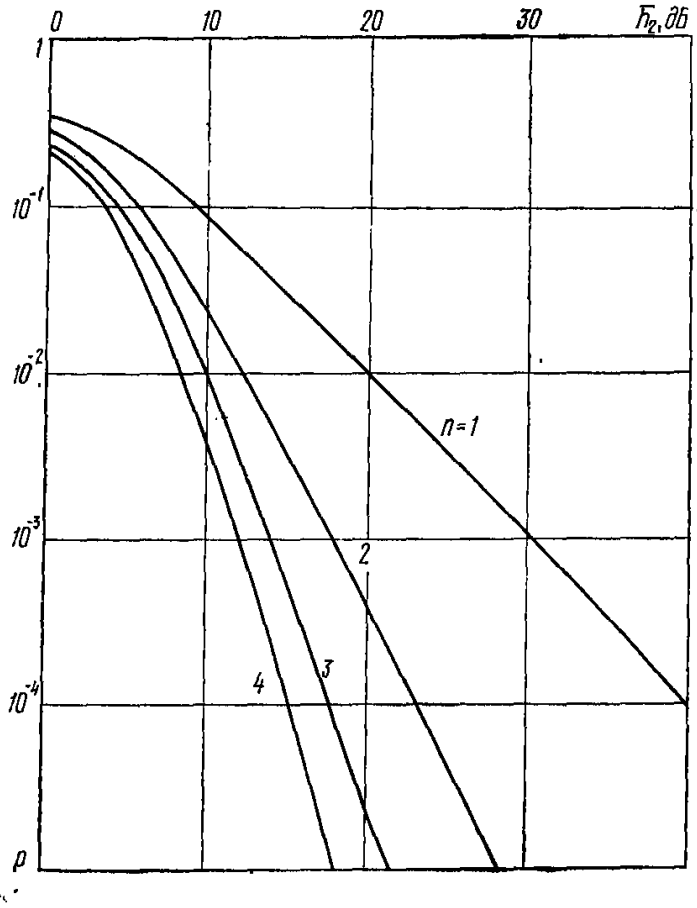


Рис. 4.24. Зависимость вероятности ошибки при разнесенном приеме в канале с рэлеевскими замираниями от $\overline{h^2}$ (двоичная система с активной паузой, ортогональная в усиленном смысле при автовыборе ветви с наиболее сильным сигналом)

росте числа ветвей. Так, при $p=10^{-4}$ переход от одиночного приема к сдвоенному дает выигрыш по мощности на 17 дБ.

Если между замираниями в отдельных ветвях имеется корреляция, выигрыш от разнесения, как и следовало ожидать, падает, однако незначительно, вплоть до коэффициента взаимной корреляции $R\approx 0{,}6$.

Можно показать, что в области малых ошибок рассмотренная схема с автовыбором незначительно уступает по помехоустойчивости схемам оптимального сложения сигналов, реализуемым более сложно.

4.8. ПРИЕМ ДИСКРЕТНЫХ СООБЩЕНИЙ В КАНАЛАХ С СОСРЕДОТОЧЕННЫМИ И ИМПУЛЬСНЫМИ ПОМЕХАМИ

Выше помехоустойчивость дискретных систем связи определялась с учетом неизбежного аддитивного флуктуационного гауссовского шума. Однако на практике приходится учитывать и действие в канале иных аддитивных помех, порождаемых внешними источниками и, прежде всего, относящихся к классу сосредоточенных по спектру («гармонических») и импульсных. В принципе, такие помехи в системах связи не являются неизбежными, что позволяет осуществлять ряд мер, уменьшающих вероятность возникновения такой помехи на входе приемного устройства или устраняющих их воздействие на демодулятор.

Заметим, что если на вход приемного устройства поступает большое число слабо коррелированных помех от различных источников сравнимой мощности, то их сумма, согласно центральной предельной теореме, представляет процесс, близкий к гауссовскому. Прибавляясь к флуктуационному шуму аппаратуры, он может существенно увеличить спектральную плотность гауссовской помехи, что потребует соответствующего увеличения мощности сигнала. Однако нередко среди множества маломощных помех на вход приемника поступают отдельные мощные импульсные или сосредоточенные помехи. В таком канале, если не принимать специальных мер, прием дискретных сообщений сопровождается дополнительными ошибками и связь может быть полностью нарушена.

Все мероприятия по защите от внешних помех можно разбить на три группы.

К первой из них относятся те, которые направлены на подавление помех в месте их возникновения, в частности, экранирование источников промышленных помех, применение искрогасящих конденсаторов, снижение уровня побочных излучений радиопередатчиков и т. п. Эти мероприятия регулируются специальными законоположениями и стандартами.

Ко второй группе можно отнести мероприятия, целью которых является воспрепятствовать проникновению помех на вход демодулятора. С этой целью в системах проводной связи осуществляется скрещивание проводов воздушных линий, совершенствуется конструкция кабелей для уменьшения взаимных влияний между жилами и улучшения экранирования от внешних влияний и т. д. В радиосвязи для этого устанавливается рациональное распределение частот между отдельными службами и каналами, с учетом размеще-

ния передатчиков и приемников и условий распространения радиоволи. Для выполнения мероприятий первых двух групп созданы международные органы, вырабатывающие допустимые нормы и контролирующие их соблюдение — международные консультативные комитеты по телеграфии и телефонии (МККТТ) и по радиосвязи (МККР).

Третья группа мероприятий, непосредственно относящаяся к данному курсу, охватывает выбор ансамбля сигналов и построения приемного устройства с целью предупредить попадание внешних помех непосредственно в решающее устройство (демодулятор) и минимизировать вероятность вызванных ими ошибок, если они все же проникнут в него.

Сосредоточенные помехи наблюдаются почти исключительно в радиоканалах. Защита демодулятора от попадания сосредоточенной помехи осуществляется обычно линейными цепями специальных блоков (входные избирательные цепи, преобразователи частоты, резонансные и полосовые усилители и т. п.) различных приемных устройств. Способность ослабить сосредоточенную помеху на входе решающей схемы приемника определяет его избирательность. Частотная избирательность обеспечивается тем, что до подачи сигнала на вход демодулятора он фильтруется в упомянутых выше линейных цепях, полоса пропускания которых достаточна для того, чтобы сигнал прошел без существенных искажений, а сосредоточенные помехи, лежащие вне полосы пропускания, при этом подавляются. Помимо частотной избирательности широко используется также пространственная избирательность, основанная на применении узконаправленных приемных антенн. Важно отметить, что воздействие сосредоточенных помех возрастает при увеличении нелинейности входных каскадов приемника, поскольку возникающие при этом комбинационные частоты (даже если помеха на входе приемника непосредственно и не попала в полосу пропускания) могут оказаться в полезной области частот. Вопросы защиты радиоприемника от сосредоточенных помех составляют основное содержание курса радиоприемных устройств.

Очевидно, что для уменьшения вероятности попадания сосредоточенной помехи в полосу частот спектра сигнала желательно использовать как можно более узкополосные сигналы. Именно поэтому в течение многих десятилетий для передачи дискретных сообщений по радио применялись только простые узкополосные сигналы (АМ, ЧМ, ОФМ), элементы которых являются отрезками синусоиды. Однако за последние 20—25 лет появилась тенденция к существенному расширению спектра сигнала путем усложнения его формы либо просто путем сокращения длительности посылки. Как уже упоминалось в предыдущем параграфе, широкополосные сигналы позволяют успешно передавать информацию в многолучевых каналах. Но, как это ни парадоксально, применение широкополосных сигналов оказалось полезным и для защиты от узкополосных сосредоточенных помех. Дело в том, что если спектр узкополосного сигнала перекрывается мощной сосредоточенной помехой, то практически не удается избежать возникновения ошибок. Если же такая помеха окажется в полосе широкополосного сигнала, то, в принципе, существует возможность «вырезать» ее режекторным фильтром (или другими способами) и по оставшейся части спектра широкополосного сигнала восстановить переданную информацию. Поэтому, хотя вероятность *попадания* сосредоточенной помехи в спектр широкополосного сигнала

больше, чем в спектр узкополосного, вероятность *ошибок*, создаваемых такой помехой, при широкополосном сигнале (и рациональио построенном приемнике) может оказаться значительно меньше.

Простейшим способом построения широкополосного сигнала для защиты от сосредоточенных помех является объединение нескольких узкополосных сигналов, передающих одинаковую информацию, на смежных полосах частот с осуществлением частотноразнесенного приема. При этом схема автовыбора строится так, что к решающему устройству подключаются только ветви, не пораженные сосредоточенными помехами.

Более сложные системы часто строятся с использованием «блока защиты от сосредоточенных помех», который представляет собой ряд параллельно включенных узкополосных фильтров со смежными полосами пропускания, рассчитанных так, что вместе они пропускают без существенных искажений весь широкополосный сигнал. Этот блок включается на вход демодулятора и управляется устройством, анализирующим напряжение на выходе каждого фильтра и запирающим те из них, в которых обнаруживаются мощные сосредоточенные помехи.

Для защиты от импульсных помех предложены различные способы, наиболее эффективные из которых основаны на амплитудном ограничении входного сигнала до его фильтрации или на мгновенном запирании приемника на время действия помехи.

В 1946 г. академик А. Н. Щукин показал, что, применяя ограничитель в широкополосном тракте приемника и пропуская ограниченный сигнал через узкополосный фильтр, можно, при надлежащем выборе полос пропускания, подавить импульсные помехи без заметного ухудшения помехоустойчивости относительно сосредоточенных и флуктуационных помех. Такая система получила название ШОУ (широкополосный фильтр, ограничитель, узкополосный фильтр). В современных устройствах роль узкополосного фильтра выполняют обычно согласованные фильтры демодулятора.

Пусть входной сигнал приемника подается на двусторонний амплитудный ограничитель с амплитудной характеристикой рис. 4.25. Если уровень U_0 выбран несколько выше напряжения полезного сигнала, то при отсутствии импульсной помехи схема приемника остается линейной. Если же появляется импульсная помеха с уровнем, бо́льшим, чем U_0, она будет ограничиваться. Таким образом, импульсная помеха длительностью $\tau_и$ со сколь угодно большой амплитудой на входе трансформируется в импульс с площадью $\tau_и U_0$. Амплитуда этого импульса примерно равна амплитуде сигнала, а спектр его сильно отличается от спектра сигнала. Поэтому после прохождения через узкополосный (или согласованный) фильтр подавляющая часть энергии импульсной помехи отсеивается и она не вызывает ошибок.

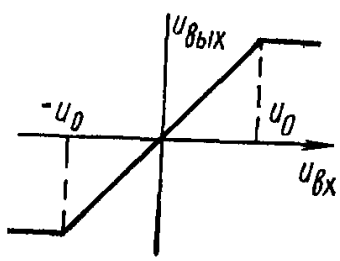


Рис. 4.25. Характеристика идеального двустороннего ограничителя

Однако в реальных условиях уровень U_0 достигается и сосредоточенной помехой, а из-за нелинейного элемента в схеме (ограничителя) образуются комбинационные частоты сосредоточенной помехи, которые в дальнейшем трудно отфильтровать. Установка ограничителя после узкополосного фильтра, устраняющего влияние сосредоточенной помехи, неэффективна, ибо на выходе такого фильтра напряжение импульсной помехи расплывается во времени и условие $\tau_и \ll T$ не может быть выполнено.

Метод мгновенного запирания приемника на время действия импульсной помехи также не лишен недостатков. Во-первых, во время запирания и отпирания возникают переходные процессы, искажающие работу демодулятора, во-вторых, суммарное входное колебание (сигнал плюс сосредоточенная и флуктуационная

помехи) оказывается прн этом промодулированным импульсом запирания, из-за чего появляются дополиительиые частотные составляющие, которые могут попасть в полосу сигнала.

Можно отметить частотно-времениую дуальность между гармонической и импульсной помехами (спектральные характеристики сосредоточенной по спектру помехи напоминают временные характеристики импульсной и наоборот). Это обстоятельство объясняет, почему меры борьбы с импульсиой и сосредоточеииой помехами в приемном устройстве взаимио противоположны.

Эффективной мерой защиты от сосредоточенных и импульсных помех является разнесенный прием одновременно по частоте и времени. Из ветвей частотного разнесения следует выбирать те, в которых меньше (или нет совсем) сосредоточенных помех, а из ветвей разнесения во времени те, где нет импульсной помехи. Весьма эффективны также методы защиты от различных помех, основанные на помехоустойчивом кодировании, которое рассматривается в гл. 5.

ГЛАВА ПЯТАЯ

ОСНОВЫ ТЕОРИИ КОДИРОВАНИЯ

5.1. НАЗНАЧЕНИЕ И КЛАССИФИКАЦИЯ КОДОВ

В этой главе рассматривается кодирование сообщений, передаваемых в дискретном канале, или кодирование в узком смысле[1]. Дискретный канал образуется из непрерывного путем включения в канал модема. На вход модулятора и с выхода демодулятора поступают дискретные кодовые символы (например, в форме импульсов), одинаковые или различные. Будем обозначать кодовые символы числами $0, 1, \ldots, m-1$, где m — основание кода.

Пусть источник выдает некоторое дискретное сообщение **a**, которое можно рассматривать как последовательность элементарных сообщений $a_i (i=1, 2, \ldots, l)$. Эти элементарные сообщения будем называть знаками, а их совокупность $\{a_i\}$ — алфавитом источника. Кодирование заключается в том, что последовательность знаков источника **a** заменяется кодовым словом, т. е. последовательностью **b** кодовых символов. Такое преобразование сообщения в кодовое слово (если не учитывать воздействия помех), как правило, является взаимно однозначным, что и позволяет осуществить декодирование, т. е. восстановить сообщение по принятому кодовому слову.

В простейшем случае, когда объем алфавита источника l равен основанию кода m, можно сопоставить каждый кодовый символ букве источника. Такое кодирование применяют, например, в морском флоте при сигнализации флагами различной формы и

[1] Как уже отмечалось, в широком смысле кодированием называют любое преобразоваиие сообщеиия в сигнал путем установления взаимного соответствия.

цвета. Чаще применяют более сложные коды, основное назначение которых заключается в согласовании источника сообщений с дискретным каналом по объему алфавита и по избыточности.

Согласование по объему необходимо во всех случаях, когда объем алфавита источника l не совпадает с количеством различных символов m, для передачи которых пригоден используемый дискретный канал. Чаще всего $l>m$, так что каждый знак источника кодируется несколькими последовательными кодовыми символами. Так, например, в простейшем телеграфном коде Бодо каждая буква русского алфавита кодируется кодовым словом из пяти двоичных символов (0 и 1); в телеграфном коде Морзе на каждую букву алфавита затрачивается от двух до шести символов, принимающих значения «точка», «тире» и «пробел».

Остановимся подробнее на согласовании источника с каналом по избыточности. Пусть случайное сообщение **A** заменяется кодовой последовательностью **B**. Поскольку считаем кодирование обратимым, то, в соответствии с (2.141) и (2.142),

$$I(\mathbf{A}, \mathbf{B}) = H(\mathbf{A}) = H(\mathbf{B}), \tag{5.1}$$

где $I(\mathbf{A}, \mathbf{B})$ — количество информации в кодовой последовательности относительно сообщения; $H(\mathbf{A})$ — энтропия сообщения; $H(\mathbf{B})$ — энтропия кодовой последовательности. Следовательно, энтропия при кодировании не изменяется[1].

Иначе обстоит дело с избыточностью, определяющей соотношение между энтропией и ее максимальным значением (при данном алфавите). Избыточность может при кодировании как возрастать, так и уменьшаться.

Пусть, например, избыточность источника велика, т. е. $H(A) \ll H_{\text{макс}}(A) = \log l$. Тогда может стоять задача о таком кодировании, при котором избыточность уменьшается (в предельном случае вовсе устраняется). Эта задача *эффективного* (или экономного) *кодирования* уже рассматривалась в § 2.10. Там было показано, что оно позволяет увеличить скорость передачи сообщений по каналу с ограниченной пропускной способностью. В частности, осмысленный русский текст можно передавать, затрачивая всего лишь 1,5 двоичных символа на букву, вместо пяти при примитивном равномерном коде.

Не будем возвращаться к методам эффективного кодирования. Отметим только некоторые свойства кодовой последовательности,

---

[1] Следует обратить внимание на то, что здесь речь идет об энтропии сообщения источника $H(A)$ и энтропии соответствующей ему последовательности кодовых символов $H(\mathbf{B})$, а не об энтропии на один символ, которая, конечно, изменяется при изменении объема алфавита. Так, например, если алфавит источника содержит 32 буквы, передаваемые равновероятно и независимо, то $H(A) = \log 32 = 5$ бит. Закодировав их двоичным кодом ($m=2$), можно представить каждую букву источника кодовым словом из пяти символов. Энтропия на каждый кодовый символ $H(B) = 1$ бит, а на все кодовое слово $H(\mathbf{B}) = 5$ бит $= H(A)$. При кодировании в реальном масштабе времени (т. е. без нарастающих задержек) справедливо равенство производительностей $H'(A) = H'(B)$ (см. § 2.10).

в которой полностью устранена избыточность. В любом месте такой последовательности все символы появляются равновероятно и независимо от значений других символов. В противном случае энтропия на символ последовательности не имела бы максимального значения $\log m$, т. е. существовала бы остаточная избыточность. Отсюда следует, что и все последовательности символов произвольно заданной длины n равновероятны. Предположим, что при передаче такой кодовой последовательности под воздействием помех возникли ошибки. Принятая ошибочная последовательность кодовых символов соответствует ошибочной последовательности сообщений, которая, однако, имеет ту же вероятность, что и правильная. Никаких признаков ошибочности принятая последовательность не может иметь. При передаче безызбыточных сигналов по каналу с ошибками любая принятая последовательность соответствует возможному сообщению, но полной уверенности в том, что именно это сообщение передано в действительности, у получателя нет. Ошибочный прием всего лишь одного кодового символа может изменить до неузнаваемости переданное сообщение. Поэтому эффективное кодирование используется в чистом виде только тогда, когда кодовая последовательность не подвергается воздействию помех.

Избыточность в передаваемом сообщении позволяет в некоторых случаях обнаруживать и исправлять ошибки. Искаженная кодовая последовательность может иметь нулевую или очень близкую к нулю вероятность, что указывает на наличие ошибки. Если определить, какая из возможных переданных последовательностей наиболее правдоподобна, можно во многих случаях ошибки исправить. Именно так читатель исправляет опечатки в книге «по контексту», а получатель телеграмм догадывается о ее подлинном содержании даже при нескольких ошибочно переданных буквах[1].

Если при кодировании не устранять, а наоборот, вводить избыточность, то должны увеличиться возможности обнаружения и исправления ошибок. Такое кодирование называется *помехоустойчивым*, или *корректирующим*. Ему посвящена основная часть этой главы.

При помехоустойчивом кодировании чаще всего считают, что избыточность источника на входе кодера $\varkappa=0$. Для этого имеются следующие основания: во-первых, очень многие дискретные источники (например, информация на выходе ЭВМ) обладают малой избыточностью; во-вторых, если избыточность первичных источников существенна, она обычно порождается сложными связями, которые в месте приема трудно использовать для повышения верности. Разумно поэтому в таких случаях по возможности уменьшить избыточность первичного источника путем эффектив-

---

[1] Именно необходимость разговаривать при воздействии акустических помех явилась причиной того, что все национальные языки в процессе своего возникновения и развития оказались избыточными и величина избыточности для всех языков близка к $\varkappa\approx 0{,}7\div 0{,}9$.

ного кодирования, а затем методами помехоустойчивого кодирования внести такую избыточность в сигнал, которая позволит достаточно простыми средствами поднять верность. Из сказанного видно, что экономное кодирование вполне может сочетаться с помехоустойчивым.

Теория кодирования за последние тридцать лет развивалась весьма интенсивно на основе современных математических методов. В настоящей книге затронуты лишь общие принципы теории кодирования. Вопросы построения используемых на практике кодов, а также технической реализации кодирующих и декодирующих устройств рассматриваются в специальных курсах.

Коды можно классифицировать по различным признакам. Одним из них является основание кода m, или число различных используемых в нем символов. Наиболее простыми являются *двоичные* (бинарные) *коды*, у которых $m=2$.

Далее коды можно разделить на *блочные* и *непрерывные*. Блочными называют коды, в которых последовательность элементарных сообщений источника разбивается на отрезки и каждый из них преобразуется в определенную последовательность (блок) кодовых символов $\{b_i\}$, называемую иногда кодовой комбинацией $\mathbf{b}_l$ $(l=1, 2, 3, \ldots, M)$. Непрерывные коды образуют последовательность символов $\{b_i\}$, не разделяемую на последовательные кодовые комбинации: здесь в процессе кодирования символы определяются всей последовательностью элементов сообщения.

В настоящее время на практике чаще всего используются блочные коды, равномерные и неравномерные. В равномерных кодах, в отличие от неравномерных, все кодовые комбинации содержат одинаковое число символов (разрядов), передаваемых по каналу элементами сигнала неизменной длительности. Это обстоятельство существенно упрощает технику передачи и приема сообщений и повышает помехоустойчивость системы синхронизации. Число различных блоков M n-разрядного равномерного кода с основанием m удовлетворяет очевидному неравенству

$$M \leqslant m^n. \tag{5.2}$$

Если в (5.2) имеет место равенство, т. е. все возможные кодовые комбинации используются для передачи сообщений, то в этом случае код называется простым, или примитивным. Он не вносит избыточности и поэтому не является помехоустойчивым.

Избыточностью равномерного кода $\varkappa_k$ называют, по аналогии с (2.127), величину

$$\varkappa_k = 1 - \frac{\log M}{n \log m}, \tag{5.3}$$

а относительной *скоростью кода*

$$R_k = \frac{\log M}{n} = (1-\varkappa_k)\log m, \ \frac{\text{бит}}{\text{симв}}. \tag{5.4}$$

Если все блоки равномерного кода передавать равновероятно и независимо друг от друга, то $\log M$ представляет собственную информацию (энтропию), приходящуюся на каждый блок, и, следовательно, R_k — это собственная информация, приходящаяся на кодовый символ.

В дальнейшем будем рассматривать, главным образом, двоичные коды ($m=2$). Напомним, что множество всех возможных двоичных блоков или *кодовых векторов* длины n образует линейное пространство (см. § 2.6), если под операцией сложения понимать поразрядное сложение по модулю 2, норму определить формулой (2.104), а под *расстоянием* понимать расстояние Хэмминга (2.105)[1].

Напомним, что расстоянием Хэмминга между двумя кодовыми n-последовательностями, b_i и b_j, которое будем далее обозначать $d(i; j)$, является число разрядов, в которых символы этих последовательностей не совпадают.

5.2. ПРИНЦИПЫ ПОМЕХОУСТОЙЧИВОГО КОДИРОВАНИЯ

Покажем, каким образом избыточное кодирование позволяет повысить верность передачи сообщения. Как уже отмечалось, для помехоустойчивых блочных равномерных кодов $m^n > M$. Это значит, что для передачи знаков сообщения используется лишь часть возможных последовательностей, составленных из m-ичных символов (часть пространства n-последовательностей). Последовательности, используемые при кодировании, называются разрешенными кодовыми комбинациями, а все другие n-последовательности — запрещенными. На вход канала поступают только разрешенные комбинации. Если при передаче кодовой комбинации $\mathbf{b}_i$ помехи не вызовут ошибок, то на выходе канала возникает та же разрешенная комбинация. Если же один или несколько символов принимается ошибочно, то на выходе канала может возникнуть одна из запрещенных комбинаций.

Таким образом, если комбинация на выходе канала оказывается запрещенной, то это указывает на то, что при передаче возникла ошибка. Отсюда видно, что избыточный код дает возможность обнаружить, в каких принятых кодовых комбинациях имеются ошибочные символы. Безусловно, не все ошибки могут быть обнаружены. Существует вероятность того, что, несмотря на возникшие ошибки, принятая последовательность кодовых символов окажется разрешенной комбинацией (но не той, которая передавалась). Однако при разумном выборе кода вероятность необнаруженной ошибки (т. е. ошибки, которая переводит разрешенную комбинацию в другую разрешенную комбинацию) может быть сделана очень малой.

---

[1] Аналогичные дискретные линейные пространства можно построить и для описания недвоичных кодов ($m>2$).

Если принята запрещенная кодовая комбинация $\hat{\mathbf{b}}_j$, то, зная параметры канала, можно определить, какая из разрешенных комбинаций $\mathbf{b}_i$ вероятнее всего передавалась, и произвести декодирование принятой комбинации $\hat{\mathbf{b}}_j$ в комбинацию, совпадающую с $\mathbf{b}_i$. Если действительно передавалась $\mathbf{b}_i$, то тем самым возникшие ошибки будут исправлены. Конечно, возможны случаи, когда в действительности передавалась не наиболее вероятная комбинация $\mathbf{b}_i$, а какая-то другая, так что декодирование окажется неправильным. Тем не менее при достаточной избыточности кода и хорошей его структуре вероятность неисправленной ошибки может быть достаточно малой и во всяком случае значительно меньшей, чем при примитивном кодировании.

Из сказанного видно, что при избыточном кодировании возможны два основных метода декодирования — с обнаружением ошибок и с исправлением ошибок[1]. Сущность метода декодирования с исправлением ошибок заключается в том, что все множество $\mathbf{B}$ принимаемых последовательностей длины n разбивается на M неперекрывающихся подмножеств: $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_M$. Если принята последовательность, принадлежащая подмножеству $\mathbf{B}_i$, то считается, что передавалась кодовая комбинация $\mathbf{b}_i$. Естественно, что в подмножество $\mathbf{B}_i$ следует включить те запрещенные комбинации $\mathbf{b}_j$, при приеме которых наиболее вероятной переданной комбинацией является $\mathbf{b}_i$.

При декодировании с обнаружением ошибок множество $\mathbf{B}$ разбивается на $M+1$ подмножеств, из которых $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_M$ содержат каждое по одной (разрешенной) кодовой комбинации, а подмножество $\mathbf{B}_{M+1}$ — все остальные (запрещенные) комбинации. В некоторых системах связи принятая запрещенная комбинация просто отбрасывается и не поступает к получателю. Это обосновано в тех случаях, когда потеря переданного сообщения значительно менее опасна, чем получение ложного сообщения. Чаще при декодировании с обнаружением ошибки ошибочно принятая кодовая комбинация не теряется, а восстанавливается специальными методами. Среди них наибольшее распространение получил метод переспроса (см. § 5.6).

Необходимо отметить, что правило декодирования с обнаружением ошибок однозначно определяется кодом (т. е. выбором разрешенных комбинаций) и не зависит от свойств канала. При исправлении ошибок, наоборот, возможны различные правила декодирования, поскольку каждую из запрещенных комбинаций можно включить в любое из подмножеств $\mathbf{B}_i$. В зависимости от свойств канала то или иное правило является предпочтительным.

Существуют и смешанные методы декодирования, когда некоторые ошибки исправляются, а некоторые только обнаруживают-

[1] Иногда говорят, что помехоустойчивые коды подразделяются на коды, обнаруживающие ошибки, и коды, исправляющие ошибки. Такое разделение необосновано, поскольку один и тот же код можно использовать при обоих методах декодирования.

ся. Здесь множество $\mathbf{B}$ также разбито на $M+1$ подмножеств, но в подмножество $\mathbf{B}_1, \ldots, \mathbf{B}_M$ помимо разрешенных комбинаций входят и некоторые близкие к ним запрещенные (исправляемые), а в $\mathbf{B}_{M+1}$ — только те запрещенные комбинации, которые не могут быть достаточно надежно исправлены.

Говорят, что в канале произошла ошибка кратности Δ, если в кодовой комбинации Δ символов приняты ошибочно. Легко видеть, что кратность ошибки есть не что иное, как расстояние Хэмминга между переданной и принятой кодовыми комбинациями, или, иначе, вес вектора ошибки (см. с. 65).

Рассматривая все разрешенные кодовые комбинации и определяя кодовые расстояния между каждой парой, можно найти наименьшее из них $d_{\text{мин}} = \min d(i; j)$, где минимум берется по всем парам разрешенных комбинаций. Это минимальное кодовое расстояние является важным параметром кода. Очевидно, что для простого кода $d_{\text{мин}} = 1$.

Обнаруживающая способность кода характеризуется следующей теоремой.

Если код имеет $d_{\text{мин}} > 1$ *и используется декодирование по методу обнаружения ошибок, то все ошибки кратностью* $\Delta < d_{\text{мин}}$ *обнаруживаются. Что же касается ошибок кратностью* $\Delta \geqslant d_{\text{мин}}$, *то одни из них обнаруживаются, а другие нет.*

Для доказательства достаточно вспомнить, что кодовое расстояние между посланной и принятой комбинациями равно Δ. Следовательно, если $\Delta < d_{\text{мин}}$, принятая комбинация не может быть разрешенной, так как это противоречило бы определению $d_{\text{мин}}$. Поэтому она будет принадлежать подмножеству запрещенных комбинаций, т. е. ошибка будет обнаружена. При $\Delta \geqslant d_{\text{мин}}$ принятая комбинация может оказаться разрешенной и ошибка останется необнаруженной, но часто и в этом случае принятая комбинация оказывается запрещенной и ошибка обнаруживается.

Процесс исправления ошибок рассмотрим сначала для симметричного канала без памяти. В таком канале, описанном в § 3.4, по определению, вероятность правильного приема символа $q = 1-p$ не зависит от того, какой символ передается, а также от того, как приняты остальные символы. Вероятность того, что вместо переданного символа b_i будет принят символ $\hat{b}_j (j \neq i)$, равна $p/(m-1)$. Отсюда легко вывести, что вероятность получения на выходе канала комбинации $\hat{\mathbf{b}}_j$, если на вход подана комбинация $\mathbf{b}_i$, равна

$$P(\hat{\mathbf{b}}_j \mid \mathbf{b}_i) = [p/(m-1)]^{d(i;\, j)} (1-p)^{n-d(i;\, j)} . \qquad (5.5)$$

Это следует непосредственно из теоремы умножения вероятностей независимых событий и из того, что для перехода $\mathbf{b}_i$ в $\hat{\mathbf{b}}_j$ необходимо, чтобы на определенных $d(i; j)$ разрядах произошли определенные ошибки, а на остальных разрядах символы были приняты верно.

Таким образом, в симметричном канале без памяти $P(\hat{b}_j|b_i)$ зависит только от кодового расстояния между $\mathbf{b}_i$ и $\hat{\mathbf{b}}_j$. В случаях, когда $p<(m-1)/m$, что практически всегда выполняется, выражение (5.5) монотонно убывает с увеличением $d(i; j)$. Следовательно, вероятность принять комбинацию $\hat{\mathbf{b}}_j$ тем больше, чем меньше ее кодовое расстояние от переданной комбинации $\mathbf{b}_i$.

Аналогично тому, как в демодуляторе по приходящему искаженному сигналу определяется передававшийся кодовый символ, в декодере по искаженной последовательности символов необходимо определить действительно передававшуюся кодовую комбинацию. Разумеется, решение, принимаемое декодером, не всегда является верным. Однако можно добиваться минимума вероятности ошибочного декодирования. Если считать, что все разрешенные кодовые комбинации передаются равновероятно, то, как было показано в § 4.2, минимальную вероятность ошибки обеспечивает решение по максимуму правдоподобия. Другими словами, при декодировании запрещенной комбинации $\hat{\mathbf{b}}_j$ декодер должен выдавать ту разрешенную комбинацию $\mathbf{b}_i$, для которой вероятность перехода $P(\hat{\mathbf{b}}_j|\mathbf{b}_i)$ больше, чем $P(\hat{b}_j|b_l)$ при $l=1, 2, \ldots, M$; $l\neq i$. Это правило декодирования по максимуму правдоподобия можно записать сокращенно так:

$$\max_i P(\hat{\mathbf{b}}_j|\mathbf{b}_i). \qquad (5.6)$$

В симметричном канале без памяти это значит, что запрещенную комбинацию $\hat{\mathbf{b}}_j$ следует декодировать, как ту разрешенную комбинацию $\mathbf{b}_i$, которая находится на наименьшем расстоянии от $\mathbf{b}_j$. Иначе говоря, в подмножество $\mathbf{B}_i$ следует включить все те комбинации $\hat{\mathbf{b}}_j$, которые ближе (в смысле Хэмминга) к $\hat{\mathbf{b}}_i$, чем к любой другой разрешенной комбинации.

Такое декодирование по наименьшему расстоянию[1] является оптимальным для симметричного канала без памяти. Однако для других каналов это правило может и не быть оптимальным, т. е. не соответствовать максимуму правдоподобия.

Исправляющая способность кода при этом правиле декодирования определяется следующей теоремой.

Если код имеет $d_{\text{мин}}>2$ и используется декодирование с исправлением ошибок по наименьшему расстоянию, то все ошибки кратностью $\Delta<d_{\text{мин}}/2$ исправляются[2]. Что же касается ошибок большей кратности, то одни из них исправляются, а другие нет.

Для доказательства покажем, что в условиях теоремы (при $\Delta<d_{\text{мин}}/2$) действительно переданная комбинация $\mathbf{b}_i$ ближе (в смысле Хэмминга) к принятой комбинации $\hat{\mathbf{b}}_j$, нежели любая дру-

[1] Это правило декодирования называют также алгоритмом Хэмминга.

[2] Подчеркнем, что здесь имеется в виду строгое неравенство, т. е. случай $\Delta=d_{\text{мин}}/2$ исключается.

гая разрешенная комбинация. Предположим противное, т. е. что существует разрешенная комбинация $\mathbf{b}_k$, для которой $d(k; j) < d(i; j)$. На основании (2.92) отсюда следует, что $d(k; i) \leqslant d(k; j) + d(i; j) < 2d(i; j)$. Но по условию теоремы $d(i; j) = \Delta < d_{\text{мин}}/2$. Отсюда $d(k; i) < d_{\text{мин}}$, что противоречит определению $d_{\text{мин}}$. Это противоречие и свидетельствует о справедливости теоремы.

Две доказанные теоремы позволяют оценить вероятность ошибочного декодирования (при декодировании с исправлением ошибок) и вероятность необнаруженной ошибки (при декодировании с обнаружением ошибок) в симметричном канале без памяти. Для этого напомним, что вероятность возникновения каких-либо ошибок кратности Δ определяется известным биномиальным законом

$$p(\Delta) = C_n^{\Delta} p^{\Delta} (1-p)^{n-\Delta}, \tag{5.7}$$

Эта формула следует из того, что ошибки в таком канале являются независимыми событиями с вероятностью p.

Используя доказанные теоремы и равенство (5.7), получаем следующие оценки для вероятности ошибочного декодирования $p_{\text{ош}}$ при коррекции ошибок и для вероятности необнаруженной ошибки $p_{\text{но}}$ при обнаружении ошибок:

$$p_{\text{ош}} \leqslant \sum_{\Delta=[d_{\text{мин}}/2]}^{n} C_n^{\Delta} p^{\Delta} (1-p)^{n-\Delta}; \quad p_{\text{но}} \leqslant \sum_{\Delta=d_{\text{мин}}}^{n} C_n^{\Delta} p_i^{\Delta} (1-p)^{n-\Delta}. \tag{5.8); (5.9}$$

Здесь $[d_{\text{мин}}/2]$ обозначает наибольшую целую часть $d_{\text{мин}}/2$ [знак неравенства в (5.8) и (5.9) ставится потому, что код, вообще говоря, может исправлять некоторые ошибки кратности $d_{\text{мин}}/2$ и выше и обнаруживать ошибки кратности $d_{\text{мин}}$ и выше].

Помехоустойчивые коды можно применять и в дискретных каналах со стиранием (см. с. 102). Если в принятом блоке ошибок нет, но имеется Δ_c стертых (ие опознанных модемом) символов, то при $\Delta_c < d_{\text{мин}}$ эти стирания могут быть исправлены при декодировании по минимуму расстояния. Это вытекает из определения $d_{\text{мин}}$, так как для того, чтобы две комбинации оказались неразличимыми, необходнмо стереть не менее $d_{\text{мин}}$ символов. Можно показать также, что при декодировании по минимуму расстояния в канале с ошибками и стираниями гарантированно исправляются Δ_0 ошибок и Δ_c стираний, если $\Delta_c + 2\Delta_0 < d_{\text{мин}}$.

Неравенства (5.8) и (5.9) иллюстрируют важную роль $d_{\text{мин}}$ как основного показателя исправляющих и обнаруживающих свойств кода в симметричном канале без памяти (чем больше $d_{\text{мин}}$, тем меньше $p_{\text{ош}}$ и $p_{\text{но}}$). Поэтому задача кодирования состоит в выборе кода, обладающего максимально достижимым $d_{\text{мин}}$. Впрочем, такая формулировка задачи неполна. Увеличивая длину кода n и сохраияя число кодовых комбинаций M, можно получить сколь угодно большое значение $d_{\text{мин}}$. Проще всего это достигается повторением символов кодовых комбинаций. Но совершенно очевидно, что такое «решение» задачи не представляет интереса, так как с увеличением n уменьшается возможная скорость передачи информации от источника. Если длина кода n задана, то

можно получить любое значение $d_{\text{мин}}$ (не превышающее n), уменьшая число комбинаций M. Поэтому задачу поиска наилучшего кода (в смысле максимального $d_{\text{мин}}$) следует формулировать так: при заданных M и n найти код длины n, содержащий M комбинаций и имеющий наибольшее возможное $d_{\text{мин}}$. В общем виде эта задача в теории кодирования не решена, хотя для многих значений n и M ее решения получены.

Эффективность помехоустойчивого кода возрастает при увеличении его длины. Это вытекает из формулы (3.79), так как вероятность ошибочного декодирования уменьшается при увеличении длины кодируемого сообщения.

Будем рассматривать не один определенный код, а множество различных помехоустойчивых кодов, имеющих одинаковую относительную скорость [см. ф-лу (5.4)] $(\log M)/n = R_k$.

Расположим коды в порядке увеличения n (а следовательно, и M). Предположим, что эти коды используются в симметричном канале без памяти, вероятность ошибки в котором равна p. Математическое ожидание кратности ошибок в кодовой комбинации $\Delta = np$.

По закону больших чисел при сколь угодно малом положительном ε вероятность того, что $\left|\frac{\Delta}{n} - p\right| < \varepsilon$, стремится к 1, когда n неограниченно возрастает. Следовательно, если из нашего множества кодов выбрать код с достаточно большим n, то с вероятностью $1-\delta$ (где δ — сколь угодно малое положительное число) кратность ошибок будет лежать в пределах

$$n(p-\varepsilon) \leqslant \Delta \leqslant n(p+\varepsilon). \tag{5.10}$$

Если для выбранного кода $d_{\text{мин}} > 2n(p+\varepsilon)$, то он позволяет исправлять все ошибки кратностью, меньшей $n(p+\varepsilon)$, а значит, в соответствии с (5.10), с вероятностью $1-\delta$ все кодовые комбинации будут декодированы верно. Таким образом, для того, чтобы с увеличением n вероятность правильного декодирования в симметричном канале без памяти стремилась к единице[1], достаточно выполнить условие

$$\lim_{n\to\infty} \frac{d_{\text{мин}}}{n} > 2(p+\varepsilon). \tag{5.11}$$

В связи с этим возникает вопрос, при каких значениях p и M условие (5.11) выполняется.

При рассмотрении данного вопроса ограничимся для простоты двоичным случаем $m=2$. В теории кодирования доказывается, что

---

[1] Следует подчеркнуть, что речь идет о пределе вероятности правильного декодирования при увеличении n и M, когда относительная скорость кода $(\log M)/n$ фиксирована. Это значит, что, увеличивая длину блока n, необходимо увеличивать также длину последовательности элементарных сообщений, кодируемых этим блоком, а следовательно, и M.

при достаточно больших длинах блоков n всегда существуют коды с $d_{\text{мин}}$, удовлетворяющим неравенству

$$\frac{\log M}{n} < 1 - H\left(\frac{d_{\text{мин}}}{n}\right), \tag{5.12}$$

где $H\left(\frac{d_{\text{мин}}}{n}\right)$ — значение функции $H(x) = -x\log x - (1-x)\log(1-x)$, вычисленное при $x = d_{\text{мин}}/n$. Неравенство (5.12) играет важную роль в теории кодирования и называется границей Варшамова — Гильберта.

Очевидно, что левая часть (5.12) есть скорость двоичного кода. Поэтому из неравенств (5.11) и (5.12) следует, что всегда можно построить код, для которого вероятность правильного декодирования в двоичном симметричном канале стремится к единице, когда скорость кода удовлетворяет неравенству

$$(\log M)/n < 1 - H(2p). \tag{5.13}$$

С другой стороны, из теоремы Шеннона (см. § 3.7) следует, что существует код, позволяющий обеспечить сколь угодно малую вероятность ошибочного декодирования, если относительная скорость передачи информации $(\log M)/n$ меньше пропускной способности $C_{\text{симв}}$. В двоичном симметричном канале, согласно (3.57), $C_{\text{симв}} = 1 - H(p)$. Следовательно, оценка скорости передачи неравенством (5.13) слабее, чем даваемая теоремой Шеннона. Это объясняется тем, что неравенство (5.11) является достаточным, но не необходимым условием, а (5.12) дает лишь нижнюю границу для $d_{\text{мин}}$. Тем не менее, проведенные рассуждения об асимптотике $d_{\text{мин}}$ имеют важное значение, особенно при рассмотрении проблемы сложности декодирования.

Полученные выше результаты на первый взгляд говорят о том, что помехоустойчивое кодирование (по крайней мере, для симметричного канала без памяти) реализуется весьма просто. В память кодирующего устройства (кодера) записываются разрешенные кодовые комбинации выбранного кода и правило, по которому с каждым из M сообщений источника сопоставляется одна из таких комбинаций. Данное правило известно и на декодере.

Получив от источника определенное сообщение, кодер отыскивает соответствующую ему комбинацию и посылает в канал. В свою очередь, декодер, приняв комбинацию, искаженную помехами, сравнивает ее со всеми M комбинациями списка и отыскивает ту из них, которая ближе остальных к принятой[1]. Однако даже при умеренных значениях n такой способ оказывается весьма сложным. Покажем это на примере.

Пусть для двоичного кода выбрано значение $n = 100$, обеспечивающее приемлемую вероятность выполнения (5.10), а скорость кода $(\log M)/n$ примем равной 0,5. При этом условие (5.13) выполняется для любого двоичного симметричного канала, у которого $p < 0{,}055$. Тогда $\log M = 50$ и $M = 2^{50} \approx 10^{15}$. Таким образом, кодовая таблица должна содержать 10^{15} кодовых комбинаций, или

[1] Из этого видно, что операция декодирования значительно сложнее операции кодирования. Сказанное справедливо для всех кодов при декодировании с исправлением ошибок.

$100 \cdot 10^{15} = 10^{17}$ кодовых символов. В аппаратуре кодера и декодера эти таблицы «записываются» на двоичных запоминающих ячейках: например, магнитных сердечниках, магнитной ленте, триггерах, криотронах и т. п. Предположим, что в результате успехов микроэлектроники, через несколько лет удастся производить подобную запись, затрачивая на каждый двоичный символ объем в 0,01 мм^3, или 10^{-5} см^3. Такие запоминающие устройства в настоящее время можно найти только на страницах фантастических романов. Вся таблица в данном случае займет объем $10^{17} \cdot 10^{-5} = 10^{12}$ см^3. Это объем куба, каждая сторона которого равна 100 м. Очевидно, что изготовление такого устройства совершенно нереально[1]. Но им не исчерпываются кодер и декодер. В частности, в декодере необходимо проделать 10^{17} операций, сравнивая символы принятой комбинации с символами, хранящимися в таблице. Так как на это можно отвести только время порядка длительности кодовой комбинации (например, 1 с), а число операций, выполняемых в одну секунду электронными логическими схемами, не превышает в настоящее время 10^6—10^7, то пришлось бы применить около 10^{10} параллельно работающих схем сравнения.

Таким образом, применение достаточно эффективных (а значит, и достаточно длинных) кодов при табличном методе кодирования и декодирования технически невозможно. Поэтому основное направление теории помехоустойчивого кодирования заключается в поисках таких классов кодов, для которых кодирование и декодирование осуществляются не путем перебора таблицы, а с помощью некоторых регулярных правил, определенных алгебраической структурой кодовых комбинаций. Один из таких классов представляют *линейные* коды, которые, в свою очередь, содержат различные подклассы кодов, отличающиеся теми или иными свойствами. Некоторые из них позволяют существенно упростить построение кодера и декодера, даже при значениях $n > 100$. Оказывается выгодным применять коды с $d_{\text{мин}}$ меньше оптимального, если их структура позволяет упростить кодирование и особенно декодирование.

5.3. ЛИНЕЙНЫЕ ДВОИЧНЫЕ БЛОЧНЫЕ КОДЫ [2]

Линейными называются такие двоичные коды, в которых множество всех разрешенных блоков является линейным пространством относительно операции поразрядного сложения по модулю 2. Очевидно, что оно является подпространством линейного пространства, образуемого множеством всех (а не только разрешенных)

---

[1] Заметим, что число нейронов в нервной сети человеческого мозга по современным оценкам составляет «всего» около 10^{10}.

[2] Можно построить и недвоичные линейные коды, если основание m — простое число или степень простого числа, однако это требует более сложного алгебраического аппарата. В настоящее время чаще всего применяют двоичные коды.

двоичных последовательностей длины n. Для того чтобы разрешенные блоки были линейным пространством (см. § 2.5), они должны содержать нулевой элемент, т. е. блок, состоящий из n нулей, должен быть разрешенным, а сумма (поразрядная по модулю 2) любых разрешенных блоков должна быть также разрешенным блоком. Двоичные линейные коды называют также *групповыми*. Линейный код можно построить следующим образом. Среди всех 2^n последовательностей кодовых символов можно выбрать различными способами n линейно независимых. В частности, ими могут быть (но не обязательно) элементы, образующие ортогональный базис. Выберем из них k любых линейно независимых блоков, которые обозначим $\beta_1, \beta_2, \ldots, \beta_k$ $(k<n)$, и образуем все линейные комбинации вида

$$\mathbf{b}_i = \sum_{l=1}^{k} \alpha_{il}\,\boldsymbol{\beta}_l, \tag{5.14}$$

где α_{il} принимает значения 0 или 1, а суммирование — поразрядное по модулю 2*. Легко видеть, что множество $\{\mathbf{b}_i\}$ этих комбинаций образует линейное пространство, содержащее 2^k блоков, т. е. линейный код.

Действительно, множество $\{\mathbf{b}_i\}$ содержит нулевой элемент, получающийся, когда все $\alpha_{il}=0$ $(l=1,\ldots,k)$. Сумма любых двух элементов $\mathbf{b}_i+\mathbf{b}_j=\sum_{l=1}^{k}(\alpha_{il}+\alpha_{jl})\,\boldsymbol{\beta}_l$ представляет собой также элемент этого множества, поскольку $\alpha_{il}+\alpha_{jl}$ принимает значение 0 или 1. Число элементов множества $\{\mathbf{b}_i\}$ определяется количеством различных наборов из k двоичных коэффициентов α_{ij}, которое, очевидно, равно 2^k. Если записать k линейно независимых блоков, используемых для построения линейного кода, в виде k строк, то получится матрица размером $n\times k$, которую называют порождающей или производящей матрицей кода $\mathbf{G}$. Полученный таким образом код, содержащий 2^k блоков длиной n, обозначают (n, k). При заданных n и k существует много различных (n, k)-кодов с различными кодовыми расстояниями $d_{\text{мин}}$, определяемых различными порождающими матрицами. Все они имеют избыточность $\varkappa_k=1-k/n$, или относительную скорость $R_k=k/n$.

Заметим, что если две порождающие матрицы различаются только порядком расположения столбцов, то определяемые ими коды являются изоморфными пространствами. Такие коды называются эквивалентными. Они имеют одинаковые кодовые расстояния между соответственными парами кодовых векторов и, следовательно, одинаковые способности обнаруживать и исправлять ошибки.

Чаще всего применяются *систематические* линейные коды, которые строятся следующим образом. Сначала строится простой код длиной k, т. е. множество всех k-последовательностей двоич-

* В дальнейшем в этой главе без дополнительных оговорок будем обозначать знаками «+» и «Σ» поразрядное суммирование по модулю 2.

ных символов, называемых *информационными*. Затем к каждой из этих последовательностей приписывается $r=n-k$ *проверочных* символов, которые получаются в результате некоторых линейных операций над информационными символами. Можно показать, что для каждого линейного кода существует эквивалентный ему систематический код.

Простейший систематический код $(n, n-1)$ строится путем добавления к комбинации из $n-1$ информационных символов одного проверочного, равного сумме всех информационных символов по модулю 2. Легко видеть, что эта сумма равна нулю, если среди информационных символов содержится четное число единиц, и равна единице, если число единиц среди информационных символов нечетное. После добавления проверочного символа образуются кодовые комбинации, содержащие только четное количество единиц, т. е. комбинации с четным весом.

Такой код $(n, n-1)$ имеет $d_{\text{мин}}=2$, поскольку две различные кодовые комбинации, содержащие по четному числу единиц, не могут различаться в одном разряде. Следовательно, он позволяет обнаружить одиночные ошибки. Легко убедиться, что, применяя этот код в схеме декодирования с обнаружением ошибок, можно обнаруживать все ошибки нечетной кратности. Для этого достаточно подсчитать число единиц в принятой комбинации и проверить, является ли оно четным. Если при передаче комбинации произойдут ошибки в нечетном числе разрядов Δ, то принятая комбинация будет иметь нечетный вес и, следовательно, окажется запрещенной. Такой код называют кодом с одной проверкой на четность.

Обозначим через b_i символ кода (0 или 1), стоящий на i-м месте в кодовой комбинации. Тогда для систематического (n, k)-кода общего вида получаем следующее правило построения комбинаций $(b_1, \ldots, b_k, b_{k+1}, \ldots, b_n)$; символы $b_1, \ldots, b_k$ выбираются в соответствии с передаваемой информацией, а b_i при $i>k$ определяют так, чтобы удовлетворялись условия

$$\sum_{i=1}^{n} \gamma_{ij} b_i = 0, \quad \begin{matrix} i=1, \ldots, n, \\ j=k+1, k+2, \ldots, n, \end{matrix} \tag{5.15}$$

где γ_{ij} — коэффициенты (0 или 1), характеризующие данный код.

Если набор всех коэффициентов γ_{ij} собрать в таблицу (матрицу), то получим так называемую *проверочную матрицу* кода $\mathbf{H}$ размерности $n \times (n-k)$:

$$\mathbf{H} = \begin{pmatrix} \gamma_{1,\,k+1}, & \ldots, & \gamma_{n,\,k+1} \\ \ldots & \ldots & \ldots \\ \gamma_{1,n} & & \gamma_{n,n} \end{pmatrix}. \tag{5.16}$$

Единицы в каждой i-й строке матрицы $\mathbf{H}$ показывают, какие информационные символы нужно сложить, чтобы получить i-й проверочный символ.

Из (5.15) легко получить, что произведение порождающей и транспонированной проверочной матриц

$$\mathbf{G}\mathbf{H}' = \mathbf{0}. \tag{5.17}$$

Здесь произведение матриц, состоящих из двоичных чисел, понимается в обычном смысле, но с учетом того, что сложения производятся по модулю 2. Штрих обозначает транспонирование, а $\mathbf{0}$ — нулевую матрицу размерности $k \times (n-k)$.

Для рассмотренного примера кода $(n, n-1)$ с четным весом проверочная матрица вырождается в вектор-строку длиной n:

$$\mathbf{H} = (1, 1, 1, \ldots, 1),$$

а порождающая матрица имеет вид

$$\mathbf{G} = \begin{pmatrix} 1, 0, 0, 0, \ldots, 1 \\ 0, 1, 0, 0, \ldots, 1 \\ 0, 0, 1, 0, \ldots, 1 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ 0, 0, 0, 0, \ldots, 1, 1 \end{pmatrix}.$$

Рассмотрим другой пример систематического кода — код (7, 4), заданный порождающей матрицей

$$\mathbf{G} = \begin{pmatrix} 1, 0, 0, 0, 1, 1, 0 \\ 0, 1, 0, 0, 0, 1, 1 \\ 0, 0, 1, 0, 1, 1, 0 \\ 0, 0, 0, 1, 1, 0, 1 \end{pmatrix} \tag{5.18}$$

или проверочной матрицей

$$\mathbf{H} = \begin{pmatrix} 1, 0, 1, 1, 1, 0, 0 \\ 1, 1, 1, 0, 0, 1, 0 \\ 0, 1, 1, 1, 0, 0, 1 \end{pmatrix}. \tag{5.19}$$

Этот код имеет $d_{\text{мин}} = 3$ и позволяет обнаруживать все одиночные и двойные ошибки или исправлять (по алгоритму Хэмминга) все одиночные ошибки.

Заметим, что строки проверочной матрицы являются линейно независимыми векторами. Следовательно, проверочная матрица может служить порождающей матрицей для другого линейного кода $(n, n-k)$, называемого двойственным. Так, например, матрица (5.19) является порождающей матрицей кода (7, 3), имеющего $d_{\text{мин}} = 4$. Матрица (5.18) является проверочной для этого кода.

Преимуществом линейных, в частности систематических, кодов является то, что в кодере и декодере не нужно хранить список из 2^k блоков, а при декодировании не нужно вычислять 2^k расстояний. Вместо этого достаточно хранить, например, $n-k$ строк проверочной матрицы и при декодировании проверять выполнение $n-k$ равенств (5.15). Так, например, при коде (100, 50) нужно хранить 50 строк по 100 символов, т. е. всего 5000 символов, а не 10^{17}, как при табличном кодировании, и проверять 50 равенств, вместо перебора 10^{15} расстояний.

Обнаружение и исправление ошибок при систематическом коде можно производить следующим образом. В режиме обнаружения ошибок осуществляется проверка выполнения равенств (5.15), и если хотя бы одно из них не выполнено, принятый блок бракуется как ошибочный. В режиме исправления ошибок после проверки равенств (5.15) строится последовательность $\mathbf{c}=(c_1, c_2, \ldots, c_{n-k})$, называемая *синдромом*[1], где $c_j (j=1, \ldots, n-k)$ — двоичный символ, равный нулю, если j-е равенство в (5.15) выполнено, и единице, если оно не выполнено. Нулевой синдром указывает на то, что все проверки выполнены, т. е. принятый блок является разрешенным. Всякому ненулевому синдрому соответствует определенная конфигурация ошибок, которая и исправляется. Так, например, для рассмотренного кода (7, 4) в табл. 5.1 представлены ненулевые синдромы и соответствующие конфигурации ошибок.

Таблица 5.1

Синдром	001	010	011	100	101	110	111
Конфигурация ошибок	0000001	0000010	0100000	0000100	0001000	1000000	0010000

Таким образом, код (7, 4) позволяет исправить все одиночные ошибки.

Из этого примера видно, что при декодировании систематического кода нет надобности перебирать таблицу, содержащую 2^k кодовых комбинаций, и производить $n-k$ сравнений. Помимо тех операций, которые осуществляются в кодере, декодер должен произвести $n-k$ сравнений и перебрать таблицу исправлений, содержащую 2^{n-k} строк. Для такого кода, как (7, 4), это осуществляется просто, но зато код оказывается малоэффективным.

Как уже неоднократно отмечалось, для получения высокой верности связи следует применять коды с достаточно большой длиной. Однако с ростом n, если отношение k/n (скорость кода) фиксировано, растет и разность $n-k$, а следовательно, и объем таблицы исправлений, равный 2^{n-k}. Так, для кода (63, 45) он равен $2^{18}=262\,144$.

Следовательно, применение систематического кода в общем случае хотя и позволяет упростить декодирование по сравнению с табличным методом, все же при значениях n порядка нескольких десятков не решает задачи практической реализации.

Если (n, k)-код используется для обнаружения ошибок, то в теории кодирования доказывается, что при любой вероятности

[1] Термин «синдром» заимствован из медицины, где он обозначает сочетание симптомов, характерных для определенного заболевания. В данном случае синдром представляет сочетание результатов проверок равенств (5.15), характерное для определенной конфигурации ошибок.

ошибочного приема символа $p \leqslant 1/2$ найдется код, для которого

$$p_{\text{но}} \leqslant 1/2^{n-k}. \tag{5.20}$$

Таким образом, увеличивая число корректирующих символов, можно обеспечить сколь угодно малую вероятность необнаруженной ошибки (и, следовательно, вероятность выдачи получателю ложной информации). Однако для сохранения скорости кода k/n это потребует увеличения длины блока n. При большой длине блока n вероятность появления обнаруживаемой ошибки возрастает и, следовательно, увеличиваются трудности восполнения потерянной информации.

В последнее время проблеме декодирования уделялось большое внимание и достигнуты значительные успехи. Был предложен ряд кодов и способов декодирования, при котором сложность декодера растет не экспоненциально, а лишь как некоторая степень n. Более подробно эти способы изучаются в специальных технических курсах. Здесь же укажем лишь наиболее важные из них.

5.4. НЕКОТОРЫЕ РАЗНОВИДНОСТИ СИСТЕМАТИЧЕСКИХ КОДОВ

В поисках более простой техники кодирования и декодирования был найден подкласс линейных систематических двоичных кодов, называемых *циклическими* и нашедшими широкое применение в технике связи. Название этих кодов связано с тем, что каждый вектор, получаемый из кодового путем циклической перестановки его символов, также принадлежит коду. Примером циклического кода является код (7, 4), определяемый матрицами (5.18) и (5.19). Так, например, циклические перестановки вектора 1000101 являются также кодовыми векторами 0001011, 0010110, 0101100 т. д.

Теория циклических кодов основана на изоморфизме пространства двоичных n-последовательностей и пространства полиномов степени не выше $n-1$ вида

$$b(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_{n-1} x^{n-1}, \tag{5.21}$$

где коэффициенты α принимают значения 0 и 1. Множество таких полиномов образует линейное пространство, если определить сложение полиномов как суммирование коэффициентов при соответствующих степенях x по модулю 2. Переменная x является символической, и ее значение никак не влияет на свойства пространства.

Легко видеть, что между полиномами (5.21) и n-последовательностями имеется изоморфизм, причем полиному (5.21) соответствует n-последовательность $\alpha_{n-1}, \alpha_{n-2}, \ldots, \alpha_1, \alpha_0$.

Любой полином $g(x)$ степени $r<n$, который делит без остатка двучлен x^n-1, может служить порождающим полиномом циклического (n, k)-кода, где $k=n-r$. В этот код входят те полиномы (5.21), которые без остатка делятся на $g(x)$. В частности, для кода (7, 4) порождающим полиномом является

$$g(x) = 1 + x + x^3. \tag{5.22}$$

Среди циклических кодов особое значение имеет класс кодов, предложенных Боузом и Рой-Чоудхури и независимо от них Хоквингемом. Коды Боуза — Чоудхури — Хоквингема (обозначаемые сокращением БЧХ) отличаются специальным выбором порождающего полинома, вследствие чего для них существуют сравнительно просто реализуемые процедуры декодирования. Доказано, что для

любой пары натуральных чисел s и $\Delta < 2^{s-1}$ можно построить двоичный код БЧХ с $n=2^s-1$ и $k \geqslant n-s\Delta$, исправляющий любую конфигурацию ошибок с кратностью, меньшей или равной Δ.

Относительно простой является процедура *мажоритарного декодирования*, применимая для некоторого класса двоичных линейных, в том числе циклических кодов. Основана она на том, что в этих кодах каждый информационный символ можно несколькими способами выразить через другие символы кодовой комбинации. Если для некоторого символа эти способы проверки дают неодинаковые результаты (одни дают результат «0», а другие — «1», что может быть только в случае ошибочного приема), то окончательное решение по каждому из информационных символов принимается по мажоритарному принципу, т. е. по большинству. Мажоритарное декодирование осуществляется относительно просто, посредством регистров сдвига.

Примером кода, допускающего мажоритарное декодирование, является код (7, 3), двойственный рассмотренному ранее коду (7, 4). Для него матрица (5.19) является образующей. Ее удобнее записать, переставив столбцы, так:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Символ b_1 в этом коде связан с другими символами следующими соотношениями: $b_1=b_2+b_6=b_7+b_5=b_3+b_4$. Обозначим принятые после демодуляции символы $\hat{b}_1$, $\hat{b}_2$, $\hat{b}_3$, $\hat{b}_4$, $\hat{b}_5$, $\hat{b}_6$, $\hat{b}_7$. Если бы они все были приняты верно, то для переданного символа b_1 были бы верны следующие четыре равенства (проверки):

$$b_1=\hat{b}_1,\quad b_1=\hat{b}_2+\hat{b}_6,\quad b_1=\hat{b}_5+\hat{b}_7,\quad b_1=\hat{b}_3+\hat{b}_4. \tag{5.23}$$

Каждый из символов принятого блока входит только в одну из этих проверок для b_1. Код, для которого выполняется это условие, называется кодом с разделенными проверками.

Предположим, что один из символов блока принят с ошибкой. Тогда три из проверок (5.23) дадут одно (правильное) значение b_1, а одна проверка, в которой участвует ошибочный символ, даст другое, ошибочное значение. Принимая решение по большинству, получим правильное значение b_1. Если бы ошибок было три или больше, то мажоритарное решение могло бы дать ошибочный результат. При двух ошибочно принятых символах может получиться одинаковое (верное) значение для b_1 во всех проверках (например, при ошибочно принятых $\hat{b}_2$ и $\hat{b}_6$) либо в двух проверках $b_1=0$, а в двух других $b_1=1$ (например, при ошибочно принятых $\hat{b}_2$ и $\hat{b}_7$). В последнем случае можно только констатировать наличие ошибок.

После принятия решения о символе b_1 аналогичным образом проверяются значения b_2 и b_3. Поскольку рассматриваемый код

(7, 3) — циклический, соответствующие проверки получаются из (5.23) циклической перестановкой.

Мощные коды (т. е. коды с длинными блоками и большим кодовым расстоянием $d_{\text{мин}}$) при сравнительно простой процедуре декодирования можно строить, объединяя несколько коротких кодов. Так строится, например, *итеративный код* из двух линейных систематических кодов (n_1, k_1) и (n_2, k_2). Вначале сообщение кодируется кодом 1-й ступени (n_1, k_1). Пусть k_2 блоков кода 1-й ступени записаны в виде строк матрицы. Ее столбцы содержат по k_2 символов, которые будем считать информационными для кода 2-й ступени (n_2, k_2) и допишем к ним n_2-k_2 проверочных символов. В результате получится блок (матрица $n_1 \times n_2$), содержащий $n_1 n_2$ символов, из которых $k_1 k_2$ являются информационными:

	информационные символы 1-й ступени	проверочные символы 1-й ступени
Информационные символы 2-й ступени	$b_{1,1},\ b_{1,2},\ldots,\ b_{1,k_1}$	$b_{1,k_1+1},\ b_{1,k_1+2},\ldots,\ b_{1,n_1}$
	$b_{2,1},\ b_{2,2},\ldots,\ b_{2,k_1}$	$b_{2,k_1+1},\ b_{2,k_1+2},\ldots,\ b_{2,n_1}$
	. . . . .	. . . . .
	$b_{k_2,1},\ b_{k_2,2},\ldots,\ b_{k_2,k_1}$	$b_{k_2,k_1+1},\ b_{k_2,k_1+2},\ldots,\ b_{k_2,n_1}$
Проверочные символы 2-й ступени	$b_{k_2+1,1},\ b_{k_2+1,2},\ldots,\ b_{k_2+1,k_1}$	$b_{k_2+1,k_1+1},\ b_{k_2+1,k_1+2},\ldots,\ b_{k_2+1,n_1}$
	. . . . .	. . . . .
	$b_{n_2,1},\ b_{n_2,2},\ldots,\ b_{n_1,k_1}$	$b_{n_2,k_1+1},\ b_{n_2,k_1+2},\ldots,\ b_{n_2,n_1}$.

Процесс построения кода можно продолжить в 3-м измерении и т. д.

При декодировании каждого блока 1-й ступени производятся обнаружение и исправление ошибок. После того как принят весь двумерный блок, осуществляется вновь исправление ошибок и стираний, но уже по столбцам, кодом 2-й ступени, причем приходится исправлять только те ошибки, которые не были исправлены (или были ложно «исправлены») кодом 1-й ступени. Легко убедиться, что минимальное кодовое расстояние для двумерного итеративного кода $d_{\text{мин}}=d_1 d_2$, где d_1 и d_2 — соответственно минимальные кодовые расстояния для кодов 1 и 2-й ступеней.

На итеративный код похож *каскадный код*, но между ними имеется существенное различие. Первая ступень кодирования в каскадном коде осуществляется так же, как в итеративном. Однако после того, как сформированы k_2 строк — блоков кода (n_1, k_1) 1-й ступени (внутреннего), они не делятся на столбцы. Каждая последовательность из k_1 двоичных (информационных) символов внутреннего кода рассматривается как один символ недвоичного кода 2-й ступени (внешнего). Основание этого кода $m=2^{k_1}$. К этим символам приписывается еще n_2-k_2 проверочных символов m-ичного кода, также в виде строк длиной k_1. К каждой из этих строк приписываются двоичные проверочные символы, в соответствии с внутренним кодом (n_1, k_1). В результате образуется такая же мат-

рица, как при итеративном двумерном коде. При приеме сначала декодируются (с обнаружением или исправлением ошибок) все строки (блоки внутреннего кода), а затем декодируется блок внешнего m-ичного кода (n_2, k_2), причем исправляются ошибки и стирания, оставшиеся после декодирования внутреннего кода. В качестве внешнего кода используется обычно m-ичный код Рида — Соломона, обеспечивающий наибольшее возможное $d_{\text{мин}}$ при заданных n_2 и k_2, если $n_2 < m$. Каскадные коды во многих случаях оказываются наиболее перспективными из известных блочных помехоустойчивых кодов.

Все операции с матрицами и полиномами над конечным полем, составляющие процедуру кодирования и декодирования, осуществляются средствами вычислительной техники. Для коротких систематических кодов они довольно просты и выполняются с помощью триггеров, объединенных в цепочки (регистры сдвига) и логические схемы. С увеличением длины кода эти операции усложняются. Хотя благодаря особой алгебраической структуре кодов БЧХ сложность растет медленно, тем не менее, при использовании эффективных длинных кодов ($n \gg 100$) приходится применять в качестве декодера электронную вычислительную машину (ЭВМ), специализированную или универсальную.

Примером неблочного кода, используемого в технике связи, может служить *рекуррентный*, или цепной, код. Это непрерывный код. В простейшем его варианте информационные символы чередуются с проверочными, образуя последовательность

$$b_1,\ b_{1,2},\ b_2,\ b_{2,3},\ b_3,\ b_{3,4},\ b_4, \ldots, \tag{5.24}$$

где b_l—l-й информационный символ, принимающий значение 0 или 1 в соответствии с передаваемым сообщением, а $b_{l,\,l+1}$ — проверочный символ, определяемый уравнением

$$b_{l,\,l+1} = b_l + b_{l+1}, \tag{5.25}$$

причем сложение производится по модулю 2.

Цепной код, содержащий на n символов k информационных, часто обозначают (k/n), в частности, код (5.24) обозначают (1/2). Ошибка при приеме одного информационного символа b_l приводит к тому, что (5.25) не выполняется для двух проверочных символов: $b_{l-1,\,l}$ и $b_{l,\,l+1}$. Поэтому алгоритм декодирования можно выразить так: если (5.25) не выполняется для двух соседних проверочных символов, то следует изменить находящийся между ними информационный символ на противоположный.

Цепной код (5.24) позволяет исправлять ошибки, расположенные на любом месте, при условии, что между двумя любыми ошибочно принятыми имеются, по крайней мере, три правильно принятых символа. Это легко показать, исследуя алгоритмы проверки (5.25).

В последние годы начали находить применение предложенные довольно давно *сверточные коды*, представляющие другую разновидность рекуррентных кодов. Для них разработаны специальные алгоритмы так называемого последовательного декодирования, позволяющие эффективно исправлять ошибки при относительно небольшой сложности. Описание этих кодов и алгоритмов [4] выходит за рамки данного учебника.

Правило декодирования по наименьшему расстоянию, как уже отмечалось, обеспечивает максимальную верность декодирования только в симметричном канале без памяти. Для несимметричных каналов и каналов с памятью (с которыми очень часто приходится встречаться на практике) оптимальными могут оказаться другие правила. Так, в каналах с памятью вероятность возникновения ошибок кратностью $\Delta > 1$ зависит от того, как расположены разряды с ошибочными символами. Смежные ошибки в этих каналах могут быть более вероятны, чем ошибки, разделенные большим числом правильно принятых символов. Для таких каналов с группированием ошибок разработаны коды и методы декодирования с обнаружением или исправлением «пачек» ошибок, т. е. ошибок большой кратности, расположенных в пределах относительно небольшого числа разрядов. В частности, пачки ошибок хорошо обнаруживаются некоторыми циклическими кодами.

Для каналов с группированием ошибок часто применяют метод *перемежения* символов, или *декорреляции* ошибок. Он заключается в том, что символы, входящие в одну кодовую комбинацию, передаются не непосредственно друг за другом, а перемежаются символами других кодовых комбинаций. Если интервал между символами, входящими в одну комбинацию, сделать длиннее «памяти» (интервала корреляции) канала, то в пределах комбинации группирования ошибок не будет и можно декодировать, как в канале без памяти. Простой пример представляет дискретный канал, образованный при передаче сигналов с ОФМ. Как было указано в § 4.5, в таком канале ошибки имеют тенденцию сдваиваться, так что вероятность появления двух смежных ошибок больше, чем вероятность появления двух несмежных ошибок, и даже больше, чем вероятность одиночной ошибки. Если в таком канале объединить в один блок 1, 3, 5-й, ... символы, а в другой блок 2, 4, 6-й, ... символы, то в этих блоках ошибки будут независимы (если нет других причин, вызывающих память) и декодирование по минимуму кодового расстояния будет оптимальным.

Очень часто «память» канала оказывается очень длинной и приходится расставлять символы, связанные кодом, на большие расстояния. Это сравнительно просто достигается при использовании цепного кода (5.24), если его несколько видоизменить. С этой целью символы, входящие в уравнение (5.25), расставляются на большое расстояние и проверочные символы связаны не с соседними информационными, а с удаленными на некоторый шаг s. В этом случае вместо (5.25) имеем

$$b_{l,\ l+s} = b_l + b_{l+s}. \qquad (5.26)$$

Величина s выбирается в соответствии с длиной памяти канала.

Упомянем еще один класс блочных, равномерных *нелинейных* кодов — коды с *постоянным весом*. В этих кодах все блоки имеют одинаковое количество единиц. Если длина блока n, а вес w, то число M возможных блоков равно C^w_n. Так, например, при $n=7$ и $w=3$ $M=C^3_7=35$. Все эти коды имеют $d_{\text{мин}}=2$, так как заменив в разрешенном блоке одну единицу нулем, а один нуль единицей, получим также разрешенный блок. Поэтому такие коды мо-

гут гарантированно только обнаруживать одиночные ошибки. Однако они используются обычно в несимметричных каналах, в которых вероятность перехода 1→0 значительно больше вероятности перехода 0→1 или наоборот. В таких каналах очень маловероятно, чтобы в одном блоке были переходы обоих видов, и поэтому почти все ошибки приводят к изменению веса блока и, следовательно, обнаруживаются. В прежние годы эти коды широко использовались (а иногда используются и сейчас) в системах с управляющей обратной связью (см. § 5.6).

5.5. ЭКВИВАЛЕНТНАЯ ВЕРОЯТНОСТЬ ОШИБКИ

Вследствие большого числа избыточных кодов необходим критерий для объективного сравнения их между собой. В известной степени таким критерием для блочных кодов может служить минимальное хэммингово расстояние $d_{\text{мин}}$. Однако оно не характеризует код в полной мере.

Так, например, код с $d_{\text{мин}}=3$ (т. е. исправляющий все одиночные ошибки) при $n=7$ неэквивалентен коду с таким же $d_{\text{мин}}=3$ при $n=50$. Применение первого из этих кодов в симметричном канале без памяти при вероятности ошибки $p=0{,}01$ позволит исправить почти все ошибки, поскольку вероятность того, что из семи символов кодовой комбинации ошибочно будут приняты два или более, очень мала (приблизительно равна 0,002). Из тысячи кодовых комбинаций в среднем только две будут декодированы неправильно. При применении же второго кода около 12% всех комбинаций будет декодировано ошибочно, потому что вероятность появления двух или большего числа ошибок в комбинации из 50 символов приблизительно равна 0,12.

Для более объективного сравнения кодов и каналов используется понятие эквивалентной вероятности ошибки. Ограничимся здесь только блочными двоичными кодами, хотя это понятие можно распространить и на другие коды.

Пусть n-разрядная кодовая комбинация содержит k двоичных единиц информации. Здесь $k=\log M$, где M — число разрешенных кодовых комбинаций. Очевидно, что для систематического кода k является числом информационных символов в комбинации. Предположим, что в некотором реальном канале I данный код обеспечивает вероятность q правильного декодирования комбинации.

Рассмотрим другой воображаемый двоичный симметричный канал без памяти II, в котором ошибки происходят с вероятностью $p_э$, и предположим, что по этому каналу передаются те же k двоичных информационных символов, но при безызбыточном кодировании. Пусть, далее, величина $p_э$ такова, что вероятность правильно принять все k символов равна q, т. е. такая же, как и вероятность правильного декодирования в канале I при использовании рассматриваемого помехоустойчивого кода. Очевидно, что с точки зрения верности передачи информации каналы I и II можно считать эквивалентными. Тогда естественно назвать величину $p_э$ эквивалентной вероятностью ошибок для канала I при заданном избыточном коде.

Величину вероятности правильного декодирования комбинации, в принципе, можно всегда вычислить, зная структуру рассматриваемого кода и свойство канала I. В крайнем случае ее можно оценить экспериментально. Зная q, можно вычислить и эквивалентную вероятность ошибок $p_э$. Для этого заметим, что в канале II

$$q=(1-p_э)^k,$$

откуда

$$p_э=1-q^{1/k}. \tag{5.27}$$

Если величина q близка к единице (как это бывает в практически важных случаях), то, обозначив $1-q=p_{\text{ош}}\ll 1$, получим

$$p_э=1-(1-p_{\text{ош}})^{1/k}\approx 1-(1-p_{\text{ош}}/k)=p_{\text{ош}}/k. \tag{5.28}$$

Величина $p_{ош}$ представляет, очевидно, вероятность ошибочного декодирования кодовой комбинации.

В качестве примера оценим эквивалентную вероятность ошибки описанного выше систематического кода (7, 4) при использовании его в симметричном канале с вероятностью ошибки p. Кодовая комбинация содержит $n=7$ символов и $k=4$ двоичных единиц информации. Код позволяет исправить все одиночные ошибки и никакие другие.

Поэтому комбинация будет декодирована правильно, если все символы приняты верно либо если из семи символов один принят ошибочно. Вероятность этого равна

$$q = (1-p)^7 + C_7^1 p\,(1-p)^6 = (1-p)^7 + 7p\,(1-p)^6.$$

При $p \ll 1$, пользуясь формулой бинома Ньютона и пренебрегая членами порядка выше p^2, получим $q \approx 1-7p+21p^2+7p-42p^2 = 1-21p^2$, откуда $p_{ош} = 1-q \approx 21p^2$, и на основании (5.28) $p_э \approx 21p^2/4 = 5{,}25p^2$.

Следует подчеркнуть, что эквивалентная вероятность ошибки характеризует код не сам по себе, а только в применении к определенному каналу. Именно это необходимо для объективной характеристики кода, поскольку один и тот же код может быть хорошим для одного канала и в то же время непригодным для другого.

5.6. СИСТЕМЫ С ОБРАТНОЙ СВЯЗЬЮ

Нередко встречаются случаи, когда информация может передаваться не только от одного корреспондента к другому, но и в обратном направлении. В таких условиях появляется возможность использовать обратный поток информации для существенного повышения верности сообщений, переданных в прямом направлении. При этом не исключено, что по обоим каналам (прямому и обратному) в основном непосредственно передаются сообщения в двух направлениях («дуплексная связь») и только часть пропускной способности каждого из каналов используется для передачи дополнительных данных, предназначенных для повышения верности.

Возможны различные способы использования систем с обратной связью в дискретном канале. Обычно они подразделяются на два типа: системы с *информационной обратной связью* и системы с *управляющей обратной связью*. Системами с информационной обратной связью называются такие, в которых с приемного устройства на передающее поступает информация о том, в каком виде принято сообщение. На основании этой информации передающее устройство может вносить те или иные изменения в процесс передачи сообщения: например, повторить ошибочно принятые отрезки сообщения, изменить применяемый код (передав предварительно соответствующий условный сигнал и убедившись в том, что он принят) либо вообще прекратить передачу при плохом состоянии канала до его улучшения.

В системах с управляющей обратной связью приемное устройство на основании анализа принятого сигнала само принимает решение о необходимости повторения, изменения способа передачи, временного перерыва связи и т. д. и передает об этом приказание передающему устройству. Возможны и смешанные методы использования обратной связи, когда в некоторых случаях решение при-

нимается на приемном устройстве, а в других случаях на передающем устройстве на основании полученной по обратному каналу информации.

Простейшим по идее методом информационной обратной связи является метод полной обратной проверки и повторения (ОПП). При этом принятый сигнал полностью ретранслируется на передающее устройство, где каждая принятая кодовая комбинация сверяется с переданной. В случае их несовпадения передающее устройство передает сигнал для стирания неправильно принятой комбинации, а затем повторяет нужную комбинацию. В качестве сигнала для стирания применяется специальная кодовая комбинация, не используемая при передаче сообщения.

Функциональная схема такой системы показана на рис. 5.1. Передаваемое сообщение, закодированное примитивным кодом, посылается в канал и одновременно записывается в запоминающем устройстве (накопителе). Принятая кодовая комбинация сразу не декодируется, а запоминается в приемном накопителе и возвращается по обратному каналу на передающий конец, где она сравнивается с переданной комбинацией. Если они совпадают, то передается следующая кодовая комбинация, в противном случае — сигнал стирания.

При этом методе окончательный ошибочный прием кодовой комбинации возможен лишь тогда, когда ошибки в принятой комбинации компенсируются ошибками, возникающими в канале обратной связи. Другими словами, для того чтобы некоторый символ в переданной кодовой комбинации был окончательно принят ошибочно, необходимо и достаточно, чтобы, во-первых, произошла ошибка в прямом канале и, во-вторых, при ретрансляции произошла такая ошибка, которая изменит неправильный ретранслируемый символ на действительно переданный. Это позволяет сразу вычислить вероятность необнаруженной ошибки для такого канала:

$$p_{\text{но}} = p_1 p_2, \tag{5.29}$$

где p_1 — вероятность ошибки в прямом канале; p_2 — вероятность противоположной ошибки в канале обратной связи.

Следовательно, если p_1 и p_2 велики, то система с полной ретрансляцией дает неудовлетворительные результаты. Практически данный метод имеет смысл в тех случаях, когда канал обратной связи обеспечивает весьма высокую верность (например, при передаче сообщений на спутник с Земли), а прямой канал имеет низкую верность (например, при передаче сообщений спутника на Землю ввиду того, что мощность передатчика на спутнике мала). Существенным недостатком системы с полной ретрансляцией является большая загрузка канала обратной связи. Существуют и более сложные системы с информационной обратной связью, в которых используются помехоустойчивые коды.

Наибольшее распространение получили системы с управляющей обратной связью (УОС) при использовании избыточных кодов для обнаружения ошибок (рис. 5.2). Такие системы часто называют *системами с переспросом,* или с автоматическим запросом ошибок, или с решающей обратной связью (РОС).

В большинстве случаев это системы дуплексные, т. е. информация в них передается в обоих направлениях. В кодере передаваемое сообщение кодируется кодом, позволяющим с большой вероятностью обнаруживать возникающие в канале ошибки.

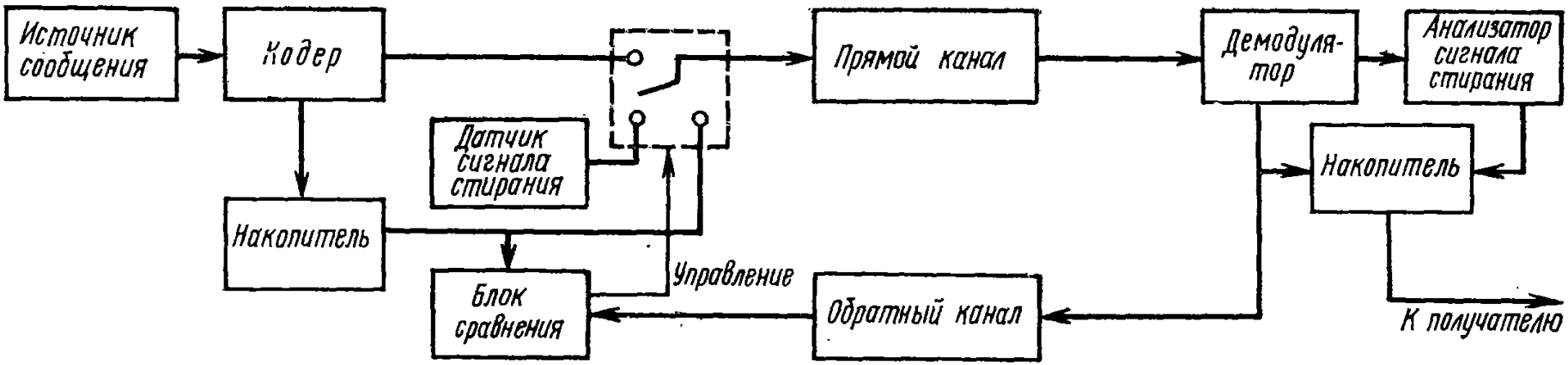


Рис. 5.1. Система с информационной обратной связью

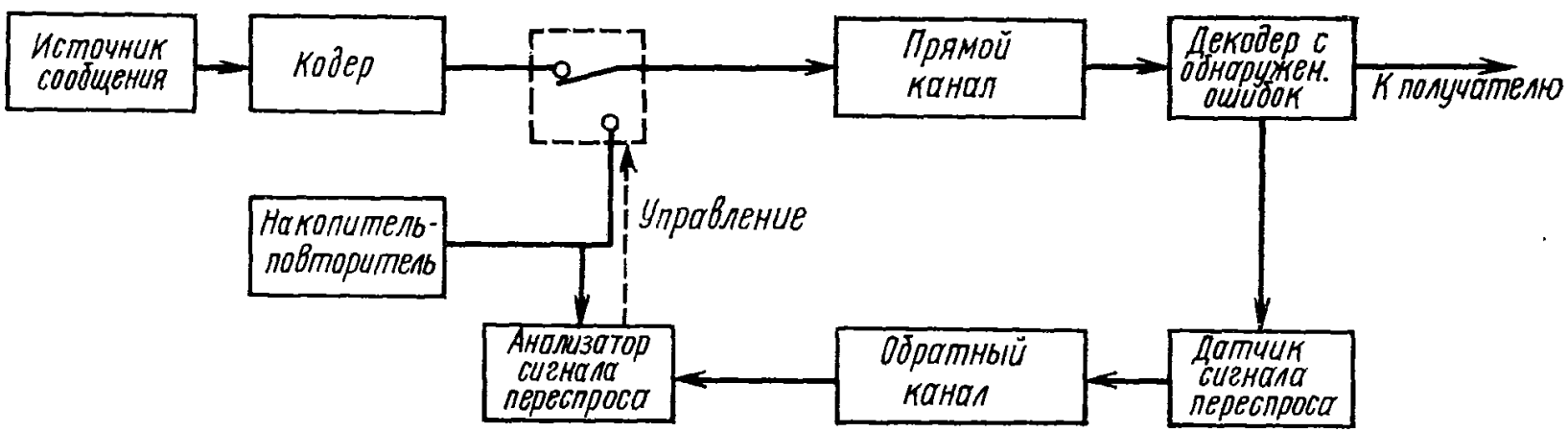


Рис. 5.2. Система с управляющей обратной связью

Принятый кодовый блок декодируется с обнаружением ошибок. Если ошибки не обнаружены, то декодированный отрезок сообщения поступает к получателю. При обнаружении ошибок блок бракуется и по обратному каналу передается специальный «сигнал переспроса». В большинстве систем этот сигнал представляет собой специальную кодовую комбинацию, на время передачи которой прерывается поток информации, идущей по обратному каналу. Прием сигнала переспроса вызывает повторение забракованного блока, который с этой целью хранится в накопителе-повторителе до тех пор, пока по обратному каналу не будет принята очередная кодовая комбинация, не содержащая переспроса. Рассмотрим несколько подробнее параметры таких систем.

Основными параметрами, характеризующими систему, являются эквивалентная вероятность ошибки и скорость передачи информации. Для их определения необходимо знать вероятности приема кодовой комбинации без ошибок $p_{пр}$, с обнаруженной ошибкой $p_{оо}$ и с необнаруженной ошибкой $p_{но}$. Эти вероятности можно вычислить, зная структуру кода и свойства канала. В частности, в симметричном канале без памяти $p_{но}$ можно оценить по формуле (5.9) или (5.20), $p_{пр}$ определить по очевидной формуле

$$p_{пр} = (1 - p)^n, \tag{5.30}$$

а $p_{оо}$ найти, исходя из того, что $p_{пр} + p_{оо} + p_{но} = 1$.

Рассмотрим передачу одной кодовой комбинации. Она может быть принята правильно с вероятностью $p_{пр}$, и тогда безошибочная информация поступает получателю; или принята с необнаруженной ошибкой (с вероятностью $p_{но}$), и тогда получателю будет выдана ошибочная информация. Наконец, она может быть принята (с вероятностью $p_{оо}$) с обнаруженной ошибкой и забракована. В этом случае, после запроса, все повторяется сначала и опять имеется возможность принять кодовую комбинацию с необнаруженной ошибкой.

Окончательно остаточная вероятность того, что кодовая комбинация будет передана получателю с необнаруженной ошибкой $p_{ост}$, складывается из вероятности необнаруженной ошибки при первой передаче, при второй передаче и т. д. Следовательно, $p_{ост} = p_{но} + p_{оо}p_{но} + p^2_{оо}p_{но} + p^3_{оо}p_{но} + \ldots$

Здесь первый член — вероятность необнаруженной ошибки при первой передаче, второй член — вероятность того, что при первой передаче возникла обнаруженная ошибка, а при повторении — необнаруженная ошибка и т. д.

Воспользовавшись формулой для геометрической прогрессии, найдем

$$p_{ост} = p_{но}\left(1 + p_{оо} + p^2_{оо} + \ldots\right) = p_{но}/(1 - p_{оо}). \tag{5.31}$$

Остаточная вероятность правильного приема $q = 1 - p_{ост}$. Отсюда можно вычислить эквивалентную вероятность ошибки. Согласно (5.27)

$$p_э = 1 - q^{1/k} = 1 - (1 - p_{ост})^{1/k} = 1 - \left(1 - \frac{p_{но}}{1 - p_{оо}}\right)^{1/k} \approx \frac{p_{но}}{k(1 - p_{оо})}. \tag{5.32}$$

Последнее приближенное равенство справедливо, если $p_{ост} \ll 1$, что на практике всегда выполняется в работоспособных системах.

Скорость передачи в системах УОС определяется не только отношением k/n, но и потерей времени на повторение[1]. Так как кодовая комбинация по-

[1] В реальных системах УОС при получении запроса приходится повторять не одну, а несколько кодовых комбинаций вследствие конечного времени распространения и обработки сигнала. Поэтому скорость оказывается меньшей, чем (5.33).

ступает к получателю только тогда, когда в ней не обнаружены ошибки, т. е. с вероятностью $1-p_{00}$, то вместо (5.4) имеем

$$R_k' = (1 - p_{00}) \log M/n,$$

или для двоичного линейного кода

$$R_k' = (1 - p_{00})\, k/n. \qquad (5.33)$$

Основным преимуществом системы УОС является простота построения декодирующего устройства. Действительно, для того, чтобы определить, принадлежит ли принятая комбинация коду, достаточно осуществить $n-k$ проверок по правилам (5.15).

Для кодов БЧХ устройства обнаружения ошибок оказываются еще более простыми и состоят из регистра длины n с обратными связями.

Система с управляющей обратной связью оказывается весьма эффективной в каналах с переменной вероятностью ошибки p (например, в коротковолновых каналах с замираниями). Когда величина p становится близкой к 1/2, т. е. пропускная способность канала падает почти до нуля, система находится в режиме постоянного переспроса, однако при хорошем коде ложная информация на выход практически не поступает. При уменьшении вероятности ошибки скорость передачи увеличивается, а верность продолжает оставаться на заданном уровне. Таким образом, система УОС как бы адаптируется (приспосабливается) к состоянию канала, используя канал настолько, насколько это оказывается возможным в каждом из его состояний.

В заключение отметим следующий факт, доказываемый в теории информации: в каналах без памяти наличие любой обратной связи не увеличивает пропускной способности прямого канала. Следовательно, если допустимо использование длинных кодов, то обратная связь не даст преимуществ. Однако, как уже указывалось, длинные коды требуют весьма сложных устройств декодирования, которые часто оказываются практически нереализуемыми. Именно в этом случае может помочь обратная связь, позволяющая реализовать ту же пропускную способность более простыми средствами.

ГЛАВА ШЕСТАЯ

ТЕОРИЯ ПЕРЕДАЧИ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

6.1. ИСТОЧНИК НЕПРЕРЫВНЫХ СООБЩЕНИЙ И ЕГО ПРОИЗВОДИТЕЛЬНОСТЬ. ВЕРНОСТЬ ПЕРЕДАЧИ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Источник непрерывных сообщений характеризуется тем, что за некоторое конечное время T он может выдать любую из бесконечного множества реализаций сообщения (сигнала). Ансамбль реализаций непрерывного источника бесконечен, а веро-

ятность появления отдельной реализации равна нулю. Если попытаться определить энтропию и производительность такого источника путем предельного перехода, как это было сделано в § 2.10 для непрерывных сигналов, то они окажутся бесконечными. Однако это не означает, что информация, интересующая получателя и подлежащая передаче по каналу, также бесконечна.

Совершенно ясно, что реальные каналы связи, а также любые устройства регистрации (в том числе и органы восприятия человека) не могут абсолютно точно передавать и регистрировать непрерывные сообщения. Практически это и не нужно, так как получатель не различает «близкие» сообщения, приводящие к одинаковым последствиям. Меру «близости» непрерывных сообщений, к сожалению, нелегко установить, ибо она в значительной степени зависит от характера сообщений и от того, как оно используется получателем.

Введем понятие об эквивалентности реализаций сообщения. Две реализации, a_1 и a_2, называются *эквивалентными*, если различие между ними несущественно в смысле определенного критерия. Так, в системе телефонной связи, если необходимо передать лишь содержание речи, то один и тот же текст, разборчиво прочитанный двумя различными дикторами (например, мужчиной и женщиной), представляет собой эквивалентные сообщения, не смотря на то, что они резко различны даже по спектру. Критерием эквивалентности сообщений здесь является разборчивость речи. При художественных вещательных передачах такой критерий не является приемлемым, ибо в этих случаях существенны и более тонкие характеристики сообщения.

В дальнейшем удобнее будет оперировать не с передаваемым непрерывным сообщением A, а с первичным сигналом $B(t)$ и его реализациями $b(t)$. Дело в том, что непрерывное сообщение A может и не быть функцией времени либо быть функцией нескольких аргументов (например, при телевизионном вещании). Первичный сигнал $B(t)$ в современных системах связи всегда является функцией времени. В тех случаях, когда и сообщение является функцией времени (например, при телефонной связи), первичный сигнал $B(t)$ точно повторяет функцию $A(t)$ и отличается от сообщения только физической природой [например, $A(t)$ — звуковое давление, $B(t)$ — ток]. Поскольку всегда преобразование сообщения в первичный сигнал обратимо, точность воспроизведения $B(t)$ предопределяет и точность воспроизведения $A(t)$. Поэтому в дальнейшем под сообщением будем понимать первичный сигнал $B(t)$, который в большинстве случаев представляет собой центрированный случайный процесс. Его дисперсию (мощность) σ_B^2 будем считать заданной.

Обеспечение необходимой верности передачи является обязательным требованием к любой системе связи. При передаче дискретных сообщений верность передачи определяется вероятностью правильного приема (или вероятностью ошибки). Такое определение верности можно распространить и на непрерывные со-

общения, если понятие «правильно» заменить понятием «эквивалентно». Тогда под верностью передачи непрерывных сообщений будем понимать вероятность того, что принятое сообщение $\hat{b}(t)$ эквивалентно переданному $b(t)$. Чтобы пользоваться таким определением верности, нужно, прежде всего, установить критерий эквивалентности. В ряде случаев этот критерий устанавливается сравнительно легко на основании требований, предъявляемых к степени точности передачи сообщений. Так, например, обстоит дело в телеметрии.

В телефонии верность передачи речевого сообщения оценивается вероятностью того, что получатель правильно поймет смысл переданного и, кроме того, узнает по голосу, кто передает сообщение. Иначе говоря, два сообщения считаются эквивалентными, если они сохраняют смысл и естественность звучания речи. Такой критерий трудно выразить аналитически, хотя практически эквивалентность реализаций устанавливается несложно. Поэтому и верность передачи чаще определяют экспериментально путем артикуляционных испытаний. Верность передачи речевых сообщений принято выражать разборчивостью фраз, т. е. процентом правильно принимаемых фраз из числа переданных. Кроме разборчивости фраз нашла применение разборчивость слов и слогов. Очевидно, критерии разборчивости слов и слогов более строгие, чем критерий разборчивости фраз.

В некоторых случаях удобно характеризовать эквивалентность сообщений по максимальной величине абсолютного значения разности

$$\delta = |\hat{b}(t) - b(t)|_{\text{макс}} \tag{6.1}$$

за время длительности сообщения T_c. Принятая реализация $\hat{b}(t)$ считается эквивалентной переданной $b(t)$, если указанная разность меньше некоторой заданной величины δ_0. Такой критерий эквивалентности является при некоторых условиях наиболее строгим и поэтому широко используется в телеметрии, где требуется особая точность восстановления переданной функции.

Наиболее универсальным методом определения эквивалентности служит критерий среднего квадрата разности между принятым и переданным сообщениями. При этом мощность передаваемого сообщения $B(t)$ считается заданной. Рассмотрим процесс

$$\varepsilon(t) = \hat{B}(t) - B(t), \tag{6.2}$$

представляющий разность между принятым и переданным случайными сообщениями. Его можно назвать *шумом воспроизведения*. В дальнейшем примем, что $\overline{\varepsilon(t)} = 0$, т. е. регулярная (систематическая) погрешность отсутствует. Сообщения будем называть эквивалентными, если среднеквадратическое отклонение $\sqrt{\overline{\varepsilon^2}}$ не превышает заданной величины ε_0:

$$\sqrt{\overline{\varepsilon^2(t)}} < \varepsilon_0, \quad \text{или} \quad \overline{\varepsilon^2(t)} < \varepsilon_0^2. \tag{6.3}$$

Взаимная информация $I(B, \hat{B})$ между двумя не тождественно равными непрерывными сообщениями (сигналами) в общем случае конечна. Будем рассматривать передаваемое сообщение $B(t)$ и множество эквивалентных ему сообщений $\hat{B}(t)$.

Очевидно, количество информации $I(B, \hat{B})$ зависит не только от статистики процесса $B(t)$, определяющей дифференциальную

энтропию $h(B)$ (см. § 2.10), но и от критерия эквивалентности, определяющего условную плотность вероятностей $w(\vec{b}|b)$, а следовательно, и условную энтропию $h(B|\hat{B})$. *Минимальное количество информации, содержащееся в сообщении $\hat{B}(t)$ относительно $B(t)$, при котором они еще эквивалентны ($\overline{\varepsilon^2(t)}<\varepsilon^2_0$), называется эпсилон-энтропией.* Согласно (2.150)

$$H_\varepsilon(B)=\min I(B,\ \hat{B})=h(B)-\max h(B\,|\,\hat{B}), \tag{6.4}$$

где минимум берется по всем условным распределениям $w(\vec{b}|b)$, для которых $\overline{\varepsilon^2(t)}<\varepsilon^2_0$. Эпсилон-энтропия определяет количество существенной информации в одном отсчете непрерывного сообщения.

Рассмотрим наиболее простой случай, когда источник непрерывного сообщения (сигнала) гауссовский, т. е. когда сообщение $B(t)$ представляет собой стационарный гауссовский процесс с заданной мощностью P_B, а критерием эквивалентности является критерий (6.3). Поскольку $B(t)=\hat{B}(t)-\varepsilon(t)$, то условная дифференциальная энтропия $h(B|\hat{B})$ при заданном сообщении $B(t)$ полностью определяется шумом воспроизведения $\varepsilon(t)$. Поэтому $\max h(B|\hat{B})=\max h(\varepsilon)$. Если шум воспроизведения $\varepsilon(t)$ имеет фиксированную дисперсию $\sigma^2_\varepsilon=\overline{\varepsilon^2(t)}$, то дифференциальная энтропия $h(\varepsilon)$ имеет максимум при нормальном распределении, равный (2.152)

$$\max h(\varepsilon)=\log\sqrt{2\pi e\sigma^2_\varepsilon}. \tag{6.5}$$

При заданной дисперсии сообщения σ^2_B дифференциальная энтропия гауссова источника $h(B)$ максимальна и равна $\log\sqrt{2\pi e\sigma^2_B}$. Следовательно, эпсилон-энтропия гауссова непрерывного источника на один отсчет

$$H_\varepsilon(B)=\log\sqrt{2\pi e\sigma^2_B}-\log\sqrt{2\pi e\sigma^2_\varepsilon}=\frac{1}{2}\log\sigma^2_B/\sigma^2_\varepsilon. \tag{6.6}$$

Величина $\sigma^2_B/\sigma^2_\varepsilon=P_B/P_\varepsilon$ характеризует минимальное отношение сигнал/шум, при котором сообщения $\hat{B}(t)$ и $B(t)$ еще эквивалентны. Обозначим это отношение ρ_0. Его значение зависит от характера передаваемых сообщений. Если отдельные отсчеты сообщения независимы, то содержащаяся в них информация складывается.

Производительность источника непрерывных сообщений можно определить как количество информации, выдаваемое источником в одну секунду при заданном критерии эквивалентности. Если источник выдает независимые отсчеты сообщения (сигнала) дискретно во времени со средней скоростью v, то его эпсилон-производительность

$$H'_\varepsilon(B)=v\,H_\varepsilon(B)=v\left[h(B)-\log\sqrt{2\pi e\sigma^2_\varepsilon}\right]. \tag{6.7}$$

Эпсилон-производительность называют также скоростью создания информации при заданном критерии верности. Для источника непрерывных сообщений, ограниченных полосой F_c, согласс-

но теореме Котельникова, шаг дискретизации $\Delta t = 1/v = 1/(2F_c)$, т. е. необходимое число отсчетов в секунду равно $2F_c$. Если спектр сообщения в полосе F_c равномерен, то эти отсчеты некоррелированны (см. пример 4 в § 2.2), а для гауссовского источника и независимы. В этом случае

$$H'_\varepsilon(B) = 2F_c H_\varepsilon(B). \tag{6.8}$$

Подставив сюда (6.6), получим для гауссовского источника с равномерным спектром в полосе F_c

$$H'_\varepsilon(B)_\Gamma = F_c \log \frac{\sigma_B^2}{\sigma_\varepsilon^2} = F_c \log \frac{P_B}{P_\varepsilon} = F_c \log \rho_0 \tag{6.9}$$

Из предыдущих рассуждений ясно, что производительность гауссовского источника квазибелого шума (6.9) больше производительности любого другого источника с той же мощностью и той же шириной спектра при том же допустимом шуме воспроизведения P_ε. Количество информации, выдаваемое гауссовским источником за время T_c,

$$T_c H'_\varepsilon(B)_\Gamma = T_c F_c \log \frac{P_B}{P_\varepsilon} = T_c F_c \log \rho_0. \tag{6.10}$$

Выражение (6.10) совпадает с характеристикой (1.1), названной в § 1.1 объемом сигнала, если $H_c = \log \rho_0$. Это означает, что объем сигнала равен максимальному количеству информации, которая содержится в сигнале длительностью T_c.

Избыточность источника непрерывных сообщений без памяти определяется соотношением

$$\varkappa_и = 1 - \frac{H_\varepsilon(B)}{H_\varepsilon(B)_\Gamma} = 1 - \frac{h(B) - \log \sqrt{2\pi e P_\varepsilon}}{(1/2)\log(P_B/P_\varepsilon)}. \tag{6.11}$$

Необходимо подчеркнуть, что шум воспроизведения, в сущности, не является аддитивной помехой. Он не всегда обусловлен шумом в канале, а является результатом различных случайных или регулярных искажений сигнала; иногда он возникает в процессе самой модуляции (см. гл. 8) и в общем случае является просто погрешностью при воспроизведении сообщения.

В соответствии с квадратичным критерием эквивалентности под верностью понимают вероятность того, что сообщение будет принято при отношении мощности сигнала к мощности шума на выходе приемника P_B/P_ε не меньше заданной величины ρ_0. Среднеквадратический критерий эквивалентности отвечает практическим запросам в большинстве случаев передачи непрерывных сообщений. Он менее строг, чем критерий максимального абсолютного отклонения, но более строг, чем большинство практических критериев, таких, как артикуляционные. Этот критерий удобен и при теоретических исследованиях.

Для непрерывного канала с пропускной способностью C, на вход которого подключен источник, обладающий производительностью $H'_\varepsilon(B)$, К. Шеннон доказал следующую теорему.

Если при заданном критерии эквивалентности сообщений источника ε^2_0 его эпсилон-производительность меньше пропускной способности канала $H'_\varepsilon(B) \leqslant C$, то существует способ кодирования и декодирования[1], *при котором неточность воспроизведения сколь угодно близка к ε^2_0. При $H'_\varepsilon(B) > C$ такого способа не существует.*

Справедливость этой теоремы вытекает из возможности дискретизации непрерывных сообщений, поскольку для дискретных сообщений аналогичная теорема доказана (см. § 3.7).

Теорема Шеннона определяет предельные возможности согласования источника непрерывных сообщений с непрерывным каналом. Оптимальное кодирование непрерывных сообщений (без дискретизации) в непрерывном канале пока не находит приемлемой реализации, хотя имеются определенные успехи в построении устройств, позволяющих существенно сократить избыточность некоторых непрерывных источников. В простейшем случае, когда канал имеет полосу пропускания, охватывающую практически весь спектр сообщения, а уровень аддитивных помех достаточно низок, используется непосредственная передача первичного сигнала $B(t)$ без модуляции. Сигнал в канале $S(t)$ при этом совпадает с $B(t)$. Примером системы связи с непосредственной передачей является телефонная связь в пределах действия одной АТС. Если на выходе канала мощность сигнала равна $P_с$, а мощность помех $P_ш$, то при $P_с/P_ш \geqslant \rho_0$ воспроизводимое сообщение будет эквивалентно переданному.

В общем же случае, как следует из теоремы Шеннона, условие $P_с/P_ш \geqslant \rho_0$ вовсе не обязательно для восстановления сообщения с заданной точностью. Для этого нужно только, чтобы пропускная способность канала превышала эпсилон-производительность источника. При этом условии можно преобразовать сообщение в сигнал так, чтобы отношение мощности P_B восстановленного сообщения к мощности шума воспроизведения (на выходе приемника) P_ε было больше ρ_0, хотя в канале (т. е. на входе приемника) отношение мощности сигнала $P_с$ к мощности шума $P_ш$ меньше (иногда во много раз), чем ρ_0. Это достигается применением помехоустойчивых видов модуляции, о которых речь будет идти в § 6.3 и далее.

6.2. НЕПОСРЕДСТВЕННАЯ ПЕРЕДАЧА СООБЩЕНИИ. ОПТИМАЛЬНАЯ ЛИНЕЙНАЯ ФИЛЬТРАЦИЯ

Пусть в канале с аддитивной помехой передается сигнал $S(t) = B(t)$, т. е. совпадающий с первичным сигналом. На выходе канала присутствует сумма сигнала и помехи:

$$Z(t) = S(t) + N(t). \tag{6.12}$$

---

[1] Здесь кодирование понимается в широком смысле, как преобразование сообщения в сигнал.

Для выявления передаваемого сообщения нужно из принятой суммы сигнала и шума $Z(t)$ выделить сигнал $S(t)$. Эту операцию называют фильтрацией. В общем случае ее можно выполнить лишь с некоторой погрешностью. Будем оценивать эту погрешность средним квадратичным отклонением оценки сигнала $\hat{S}$ на выходе фильтра в момент $t+\tau$ от входного сигнала $S(t)$:

$$\overline{\varepsilon^2(t)} = \overline{[\hat{S}(t+\tau) - S(t)]^2}. \qquad (6.13)$$

Здесь τ — некоторое заданное время запаздывания в фильтре, а среднее значение берется по ансамблям сигналов S и помех N.

Будем вначале полагать $S(t)$ и $N(t)$ стационарными взаимно некоррелированными процессами с известными спектральными плотностями $G_S(f)$ и $G_N(f)$. В общем случае фильтр, дающий оценку сигнала с минимальной среднеквадратичной погрешностью, оказывается нелинейным. Однако упростим задачу и будем отыскивать оптимальный фильтр в классе линейных цепей[1]. В такой форме задачу поставили и решили независимо друг от друга академик А. Н. Колмогоров и американский ученый Н. Винер, и поэтому оптимальный (в указанном смысле) линейный фильтр называется фильтром Колмогорова — Винера.

Обозначим искомую передаточную функцию оптимального фильтра

$$k(i2\pi f) = c(f)\, e^{i\psi(f)}, \qquad (6.14)$$

где $c(f)$ и $\psi(f)$ — действительные функции. Если бы передаточной функцией была не $k(i2\pi f)$, а

$$\tilde{k}(i2\pi f) = k(i2\pi f)\, e^{i2\pi f\tau} = c(f)\, e^{i[\psi(f)+2\pi f\tau]}, \qquad (6.15)$$

то сигнал на выходе появлялся бы на время τ раньше, т. е. не запаздывал бы по сравнению с входным сигналом. Будем сначала отыскивать передаточную функцию $\tilde{k}(i2\pi f)$ оптимального фильтра, дающего оценку сигнала без запаздывания, а затем перейдем к $k(i2\pi f)$ по формуле

$$k(i2\pi f) = \tilde{k}(i2\pi f)\, e^{-i2\pi f\tau}. \qquad (6.16)$$

Отыщем спектральную плотность $G_\varepsilon(f)$ погрешности оценки $\varepsilon(t) = \hat{S}(t) - S(t)$ для фильтра без запаздывания. Эта погрешность состоит из двух составляющих. Первая составляющая вызвана теми искажениями, которые претерпевает полезный сигнал, проходя через фильтр с передаточной функцией $\tilde{k}(i2\pi f)$. Она, очевидно, равна напряжению, которое появилось бы на выходе фильтра с передаточной функцией $\tilde{k}(i2\pi f) - 1$, если на его вход подать сигнал. Поэтому спектральная плотность мощности первой составляющей, согласно (3.7), равна $G_S(f)/\tilde{k}(i2\pi f) - 1|^2$. Вторая составляющая вызвана прохождением через фильтр помехи и имеет спектральную плотность $G_N(f)\,|\tilde{k}(i2\pi f)|^2$. Обе эти состав-

[1] Можно доказать, что применение нелинейного фильтра не приведет к уменьшению погрешности оценки, если $S(t)$ и $N(t)$ — гауссовские процессы.

ляющие между собой не коррелированы, и поэтому полная спектральная плотность погрешности $G_\varepsilon(f)$ равна их сумме:

$$G_\varepsilon(f) = G_S(f)\,|\widetilde{k}(i2\pi f) - 1|^2 + G_N(f)\,|\widetilde{k}(i2\pi f)|^2 =$$
$$= G_S(f)\{[c(f)\cos(\psi(f) + 2\pi f\tau) - 1]^2 + c^2(f)\sin^2[\psi(f) + 2\pi f\tau]\} +$$
$$+ G_N(f)\,c^2(f) = [G_S(f) + G_N(f)]\,c^2(f) + G_S(f) -$$
$$- 2G_S(f)\,c(f)\cos[\psi(f) + 2\pi f\tau]. \qquad (6.17)$$

Мощность (средний квадрат) погрешности

$$\overline{\varepsilon^2(t)} = \int_{-\infty}^{\infty} G_\varepsilon(f)\,df = \int_{-\infty}^{\infty}\{[G_S(f) + G_N(f)]\,c^2(f) +$$
$$+ G_S(f)\,[1 - 2c(f)\cos(\psi(f) + 2\pi f\tau)]\}\,df. \qquad (6.18)$$

Задача теперь сводится к выбору таких функций $c(f)$ и $\psi(f)$, которые обеспечивают минимум полученного интеграла. Что касается $\psi(f)$, то очевидно, что интеграл минимален при $\cos(\psi + 2\pi f\tau) = 1$, т. е. при $\psi = -2\pi f\tau$. Если это выполнено, то $\widetilde{k}(i2\pi f) = c(f)$ и

$$\overline{\varepsilon^2(t)} = \int_{-\infty}^{\infty}\{[G_S(f) + G_N(f)]\,c^2(f) + G_S(f)\,[1 - 2c(f)]\}\,df =$$
$$= \int_{-\infty}^{\infty}\left\{\left[\sqrt{G_S(f) + G_N(f)}\,c(f) - \frac{G_S(f)}{\sqrt{G_S(f) + G_N(f)}}\right]^2 + \right.$$
$$\left. + \frac{G_S(f)\,G_N(f)}{G_S(f) + G_N(f)}\right\}df. \qquad (6.19)$$

В подынтегральном выражении только первый член зависит от $c(f)$. Так как в квадратных скобках величина вещественна, то ее квадрат не может быть отрицательным. Следовательно, наименьшее значение интеграла обеспечивается такой функцией $c(f)$, при которой этот квадрат тождественно равен нулю, откуда

$$c(f) = \widetilde{k}(i2\pi f) = \frac{G_S(f)}{G_S(f) + G_N(f)}. \qquad (6.20)$$

При этом среднеквадратичная погрешность оценки (или мощность шума на выходе фильтра) определяется из (6.19) при подстановке (6.20):

$$\overline{\varepsilon^2(t)} = \int_{-\infty}^{\infty}\frac{G_S(f)\,G_N(f)}{G_S(f) + G_N(f)}\,df. \qquad (6.21)$$

Легко заметить, что эта погрешность равна нулю только в том случае, когда спектры сигнала и помехи не пересекаются.

Итак, получено выражение для передаточной функции линейного фильтра, которая обеспечивает оценку сигнала $S(t)$ с

минимальной среднеквадратичной погрешностью без задержки. Однако фильтр с такой передаточной функцией физически неосуществим. В этом можно убедиться, если заметить, что $c(f)$ — действительная функция. Отсюда следует, что ее преобразование Фурье, т. е. импульсная реакция фильтра, будет четной функцией. Следовательно, если она отлична от нуля при положительных t, то она будет также отлична от нуля и при $t<0$, что для реального фильтра невозможно, так как реакция не может появиться до подачи импульса.

Впрочем, при другой постановке задачи фильтрация с погрешностью (6.21) оказывается вполне возможной. Для этого нужно знать не только прошлые, но и будущие значения $z(t)$. Предположим, что $z(t)$ наблюдается и записывается в течение всего времени передачи (теоретически для $-\infty<t<\infty$) и затем обрабатывается вычислительной машиной с помощью передаточной функции (6.20). Поскольку при обработке известны и «будущие» значения $z(t)$, то это вполне осуществимо. Результат выдается в виде графика или таблицы. Примерно таким образом обрабатывались сигналы, принятые с советского искусственного спутника, сфотографировавшего в 1959 г. обратную сторону Луны.

Для техники связи интерес представляют осуществимые фильтры, работающие в реальном масштабе времени, т. е. выдающие оценку $S(t)$ путем обработки реализации $z(t)$ только до текущего момента времени t или, в крайнем случае, до момента времени $t+\tau$, где τ — небольшая задержка. Именно такой фильтр называют реализуемым. Поскольку в реализуемом фильтре не могут быть использованы значения z после момента времени $t+\tau$, то очевидно, что погрешность его будет больше, чем у нереализуемого фильтра. С увеличением τ эта погрешность уменьшается.

Остановимся на определении оптимальной передаточной функции реализуемого фильтра при $\tau=0$, т. е. работающего без задержки. С этой целью попытаемся разложить нереализуемый фильтр (6.20) на несколько фильтров и выделить из него оптимальную реализуемую часть.

Начнем с факторизации знаменателя (6.20), т. е. разложения его на два комплексно сопряженных множителя:

$$\frac{1}{G_S(f)+G_N(f)} = k_1(i\,2\pi f)\,k_1^*(i\,2\pi f) = |k_1(i\,2\pi f)|^2. \qquad (6.22)$$

Уравнение (6.22) однозначно определяет модуль $k_1(i2\pi f)$. Что же касается аргумента, то его можно выбрать произвольно, в частности, так, чтобы все нули и полюса функции комплексной переменной $k_1(\alpha+i2\pi f)=k_1(p)$ лежали в левой полуплоскости. Как известно, фильтр с такой передаточной функцией является физически реализуемым и к тому же минимально фазовым.

Теперь оптимальный нереализуемый фильтр (6.20) можно представить как последовательное соединение двух фильтров с передаточными функциями $k_1(i2\pi f)$ и $k^*_1(i2\pi f)G_S(f)$. Второй из них по-прежнему нереализуем. В частности, у сомножителя $k^*_1(i2\pi f)$ все нули и полюса лежат в правой полуплоскости.

Заметим теперь, что первый фильтр является обеляющим для входного колебания $z(t)$. Действительно, спектральная плотность процесса на его выходе согласно (3.7) равна $G_z(f)\,|k_1(i2\pi f)|^2=[G_S(f)+G_N(f)]\,|k_1(i2\pi f)|^2=1$, так что на выходе первого фильтра получается белый шум со спектральной плотностью 1. Если бы второй фильтр был реализуем, то под воздействием этого

белого шума он сформировал бы наилучшую оценку сигнала $\hat{S}(t)$. Но он физически нереализуем, а это значит, что его импульсная реакция

$$g(t) = \int_{-\infty}^{\infty} k_1^*(i\,2\pi f)\, G_S(f) \exp(i\,2\pi f t)\, df \qquad (6.23)$$

не равна тождественно нулю при $t<0$. Выразим ее в форме $g(t)=g_1(t)+g_2(t)$, где

$$g_1(t) = \begin{cases} g(t) & \text{при } t \geqslant 0, \\ 0 & \text{при } t < 0; \end{cases} \qquad g_2(t) = \begin{cases} 0 & \text{при } t \geqslant 0, \\ g(t) & \text{при } t < 0. \end{cases} \qquad (6.24);\ (6.25)$$

Фильтр с импульсной реакцией $g(t)$ (6.23) можно представить в виде параллельного соединения двух фильтров с импульсными реакциями $g_1(t)$ и $g_2(t)$. Из них первый, очевидно, реализуем, так как $g_1(t)=0$ при $t<0$, а второй — нереализуем. В результате исходный нереализуемый фильтр (6.20) представлен в виде соединения трех фильтров, как показано на рис. 6.1, из которых два реализуемы. На вход параллельно включенных фильтров Б и В подан белый шум. Фильтр Б вырабатывает ту часть оценки $\hat{S}(t)$ оптимального нереализуемого фильтра, которая определяется прошлыми значениями белого шума, а фильтр В — ту часть, которая определяется будущими значениями. Но так как в белом шуме будущие значения не зависят от прошлых, то ту часть, которую должен формировать фильтр В, в принципе, невозможно получить в реализуемом фильтре и наилучшая оценка этой части равна ее математическому ожиданию, т. е. нулю.

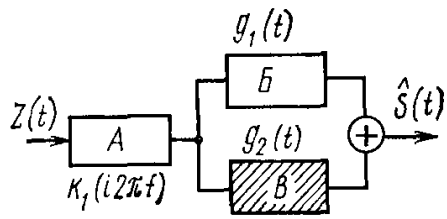


Рис. 6.1. К синтезу реализуемого фильтра Винера—Колмогорова

Таким образом, оптимальный реализуемый фильтр при $\tau=0$ получается, если из схемы рис. 6.1 удалить фильтр В, так что останутся последовательно соединенные фильтры А и Б. Передаточная функция оптимального реализуемого фильтра, очевидно, равна

$$k_p(i\,2\pi f) = k_1(i\,2\pi f) \int_0^{\infty} g_1(t) \exp(-i\,2\pi f t)\, dt =$$

$$= k_1(i\,2\pi f) \int_0^{\infty} \int_{-\infty}^{\infty} k_1^*(i\,2\pi v)\, G_S(v) \exp[i\,2\pi (v-f)\, t]\, dv\, dt. \qquad (6.26)$$

Аналогично можно синтезировать оптимальный реализуемый фильтр и для $\tau \neq 0$. При $\tau>0$ имеет место задержка и погрешность оценки с ростом τ убывает, стремясь при неограниченном увеличении τ к (6.21). При $\tau<0$ фильтр «предсказывает» будущие значения сигнала $S(t)$. Разумеется, такое предсказание имеет смысл лишь в пределах интервала корреляции сигнала и его погрешность быстро возрастает при увеличении опережения $|\tau|$.

Несколько более общую постановку задачи линейной фильтрации получим, отбросив условие стационарности сигнала $S(t)$ и помехи $N(t)$. Кроме того, учтем, что сумма сигнала и помехи (6.15) наблюдается на конечном отрезке времени.

Итак, пусть $Z(t)$ подается в некоторый момент t_0 на линейный фильтр (вообще говоря, с переменными параметрами), который должен в любой момент времени $t+\tau$ (где $t>t_0$) дать оценку $\hat{S}(t)$ с наименьшей среднеквадратичной погрешностью $\overline{\varepsilon^2(t)}$. Потребуем также, чтобы оценка была несмещенной, т. е.

чтобы математическое ожидание погрешности $\varepsilon(t)$ равнялось нулю. Известными считаются корреляционные функции сигнала $B_S(t_1, t_2)$ и помехи $B_N(t_1, t_2)$. Для отыскания структуры такого фильтра американские ученые Калман и Бьюси использовали метод переменных состояний (см. § 3.5) и получили решение в виде дифференциального уравнения, моделирующего оптимальный фильтр. При заданных начальных условиях в момент времени t_0 фильтр позволяет вырабатывать оптимальную оценку $\hat{S}(t)$ для $t > t_0$, используя каждый раз оценки, полученные для предшествующих моментов времени, и новые значения входного колебания $Z(t)$. Процедура последовательного вычисления оценок называется рекурсивной, и реализующий ее фильтр Калмана является примером оптимального рекурсивного линейного фильтра. В дальнейшем ограничимся случаем, когда $N(t)$ — белый шум (не обязательно стационарный), а задержка $\tau = 0$, и изложим основные идеи теории фильтра Калмана.

Сигнал $S(t)$ с ФК $B_S(t_1, t_2)$ можно представить как результат прохождения белого стационарного шума с единичной спектральной плотностью $V(t)$ через некоторую цепь. Полагая эту воображаемую цепь линейной (вообще говоря, с переменными параметрами), можно построить некоторое уравнение состояния

$$d\mathbf{X}/dt = \mathbf{f}(t)\,\mathbf{X}(t) + \mathbf{d}(t)\,V(t), \tag{6.27}$$

где $\mathbf{X}(t)$ — вектор состояния, первая компонента которого

$$X_1(t) = S(t). \tag{6.28}$$

Остальные компоненты, а также матрицы $\mathbf{f}(t)$ и $\mathbf{d}(t)$ определяются по ФК сигнала.

Заметим, что линейное уравнение (6.27) описывает гауссовский процесс $S(t)$, представляющий компоненту марковского процесса n-го порядка. Для описания негауссовского процесса уравнение (6.27) должно быть нелинейным. Однако, как уже отмечалось, оптимальный фильтр оказывается линейным только для гауссовских сигналов. Если же сигнал $S(t)$ негауссовский, но оптимальный фильтр ищется в классе линейных, то он оказывается таким же, как и для гауссовского процесса с той же ФК. Поэтому, не нарушая общности, можно решать задачу линейной фильтрации, полагая $S(t)$ гауссовским процессом. Для одномерного марковского процесса задача облегчается, поскольку векторные уравнения превращаются в скалярные. Для немарковского процесса можно всегда найти такое n, при котором он достаточно хорошо аппроксимируется компонентой марковского процесса n-го порядка.

Будем строить теперь другую линейную цепь — оптимальный фильтр, на вход которого подается колебание

$$Z(t) = \mathbf{c}(t)\,\mathbf{X}(t) + N(t), \tag{6.29}$$

а с выхода снимается *вектор оценок* $\hat{\mathbf{X}}(t)$, каждая составляющая которого является оценкой соответствующей составляющей вектора состояний $\mathbf{X}(t)$ (и, в частности, первая составляющая является оценкой сигнала $\hat{S}(t)$). Уравнение (6.29) является уравнением наблюдения. Оно совпадает с (6.12), если матрица $\mathbf{c}(t) = (1, 0, 0, \ldots, 0)$.

Алгоритм искомого фильтра можно представить матричным дифференциальным уравнением оценки:

$$d\hat{\mathbf{X}}/dt = \mathbf{a}(t)\,\hat{\mathbf{X}}(t) + \mathbf{g}(t)\,Z(t), \tag{6.30}$$

где $\mathbf{a}(t)$ и $\mathbf{g}(t)$ — матрицы, которые нужно найти, чтобы синтезировать фильтр.

Для нахождения $\mathbf{a}(t)$ воспользуемся условием несмещенности оценки $\overline{\boldsymbol{\varepsilon}(t)}=0$. Здесь $\boldsymbol{\varepsilon}(t)=\hat{\mathbf{X}}(t)-\mathbf{X}(t)$ — вектор ошибки. Очевидно, что при этом и $\overline{d\boldsymbol{\varepsilon}/dt}=0$. Из (6.27) и (6.30) имеем

$$d\boldsymbol{\varepsilon}/dt = d\hat{\mathbf{X}}/dt - d\mathbf{X}/dt = \mathbf{a}(t)\hat{\mathbf{X}}(t) + \mathbf{g}(t)Z(t) - \mathbf{f}(t)\mathbf{X}(t) - \mathbf{d}(t)V(t). \tag{6.31}$$

Приравняем условные математические ожидания левой и правой частей (6.31) при фиксированном входном колебании $Z(\theta)$, где $t_0 \leqslant \theta < t$. Учитывая несмещенность оценки и независимость центрированных процессов $N(t)$ и $V(t)$, а также уравнение наблюдения (6.29), получим

$$\mathbf{0} = \mathbf{a}(t)[\overline{\mathbf{X}(t)} + \overline{\boldsymbol{\varepsilon}(t)}] + \mathbf{g}(t)[\mathbf{c}(t)\overline{\mathbf{X}(t)} + \overline{N(t)}] - \mathbf{f}(t)\overline{\mathbf{X}(t)} - \mathbf{d}(t)\overline{V(t)} =$$

$$= [\mathbf{a}(t) + \mathbf{g}(t)\mathbf{c}(t) - \mathbf{f}(t)]\overline{\mathbf{X}(t)}. \tag{6.32}$$

Здесь $\overline{\mathbf{X}(t)}$ — условное математическое ожидание $\mathbf{X}(t)$ (при фиксированном $Z(\theta)$, которое не может равняться нулю при всех t. Отсюда

$$\mathbf{a}(t) = \mathbf{f}(t) - \mathbf{g}(t)\mathbf{c}(t). \tag{6.33}$$

Перепишем уравнение оценки (6.30), введя обозначение

$$\boldsymbol{\nu}(t) = Z(t) - \mathbf{c}(t)\hat{\mathbf{X}}(t) \tag{6.34}$$

и учитывая (6.33):

$$d\hat{\mathbf{X}}/dt = \mathbf{f}(t)\hat{\mathbf{X}}(t) + \mathbf{g}(t)\boldsymbol{\nu}(t). \tag{6.35}$$

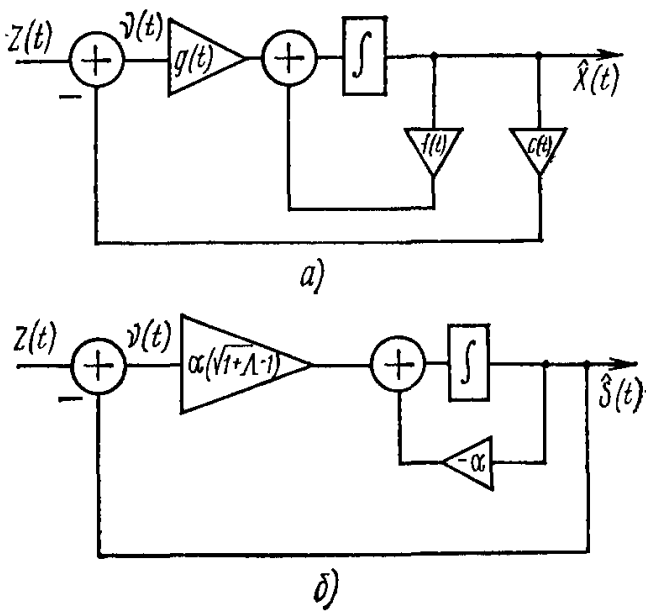


Рис. 6.2. Реализация фильтра Калмана при оптимальной оценке гауссовского сигнала

Процесс $\boldsymbol{\nu}(t)$ называют процессом обновления. Он представляет собой ту часть $Z(t)$, которую в отличие от $\hat{S}(t)$, нельзя предсказать по предшествующим наблюдениям. Можно показать, что $\boldsymbol{\nu}(t)$ является белым шумом.

Фильтр, описываемый уравнением (6.34), моделируется схемой рис. 6.2*а*. Для его построения остается определить функцию $\mathbf{g}(t)$, обеспечивающую минимум среднеквадратичной погрешности. Путем несложных, но громоздких преобразований, которые мы опускаем, можно показать, что

$$\mathbf{g}(t) = \mathbf{к}(t)G_N^{-1}(t), \tag{6.36}$$

где $G_N(t)$ — спектральная плотность белого шума $N(t)$ (вообще говоря, нестационарного, т. е. имеющего функцию корреляции $G_N(t_1)\delta(t_2-t_1)$), а вектор $\mathbf{к}(t)$ определяется матричным дифференциальным уравнением с переменными коэффициентами (уравнением Риккати):

$$d\mathbf{к}/dt = \mathbf{f}(t)\mathbf{к}'(t) + \mathbf{к}(t)\mathbf{f}'(t) - \mathbf{к}(t)G_N^{-1}(t)\mathbf{к}'(t) + \mathbf{d}(t)\mathbf{d}'(t), \tag{6.37}$$

где штрихи обозначают транспонирование матриц.

Если $\mathbf{g}(t)$ выбрать в соответствии с (6.36), то вектор средних квадратов погрешностей оценки равен

$$\overline{\boldsymbol{\varepsilon}^2(t)} = \mathbf{к}(t). \tag{6.38}$$

Таким образом, задача синтеза фильтра Калмана и оценки его погрешности решена.

Приведем простой пример. Пусть $S(t)$ — стационарный гауссовский процесс с ФК $B(t_1, t_2) = P_c \exp(-\alpha|t_2 - t_1|)$, а $N(t)$ — стационарный белый шум со спектральной плотностью $N_0/2$. Построим оптимальный фильтр Калмана для установившегося режима, т. е. положим $t_0 = -\infty$, и сравним его с фильтром Колмогорова — Винера.

Как уже отмечалось, такой процесс $S(t) = X(t)$ можно задать дифференциальным уравнением 1-го порядка[1]

$$dS/dt + \alpha S(t) = \sqrt{2\alpha P_c}\, V(t) \tag{6.39}$$

с начальным условием $S(-\infty) = 0$. В (6.39) α — постоянная, имеющая размерность частоты; $V(t)$ — центрированный белый шум с единичной спектральной плотностью. Спектральная плотность мощности сигнала выражается формулой $G_S(f) = 2\alpha P_c/(\alpha^2 + 4\pi^2 f^2)$ и шума $G_N(f) = N_0/2$.

Для синтеза фильтра воспользуемся сначала идеей рекурсивного оценивания и найдем параметры фильтра Калмана, а затем определим передаточную функцию реализуемого фильтра Колмогорова — Винера.

Сравнивая уравнения (6.39) и (6.26), видим, что в данном случае $\mathbf{f}(t) = -\alpha$; $\mathbf{d}(t) = \sqrt{2\alpha P_c}$. Так как рассматривается установившийся режим, то $\mathbf{k}(t) = \kappa$ и $d\mathbf{k}/dt = 0$. Уравнение Риккати принимает вид

$$0 = -2\alpha\mathbf{k} - \mathbf{k}^2 2/N_0 + 2\alpha P_c.$$

Это квадратное уравнение имеет два решения, из которых в силу неотрицательности дисперсии $\kappa = \overline{\varepsilon^2}$ выбираем одно, именно

$$\kappa = \alpha (N_0/2)(\sqrt{1+\Lambda} - 1), \tag{6.40}$$

где параметр $\Lambda = 4P_c/(\alpha N_0)$ характеризует отношение сигнал/шум. На основании (6.36) находим

$$\mathbf{g}(t) = g = \alpha(\sqrt{1+\Lambda} - 1). \tag{6.41}$$

В результате уравнение (6.27) примет вид

$$d\hat{S}/dt = -\alpha\hat{S}(t) + \alpha(\sqrt{1+\Lambda} - 1)[Z(t) - \hat{S}(t)]. \tag{6.42}$$

Структурная схема аналоговой модели фильтра Калмана соответствует рис. 6.2*б*. Дисперсия ошибки фильтрации определяется равенством (6.40). Решая задачу синтеза фильтра Колмогорова — Винера, получаем, согласно (6.22),

$$k_1(i2\pi f)|^2 = k_1(i2\pi f)\, k_1^*(i2\pi f) = \frac{1}{G_S(f) + \overline{G_N(f)}} =$$

$$= \frac{2(\alpha^2 + 4\pi^2 f^2)}{4\alpha P_c + N_0\alpha^2 + 4N_0\pi^2 f^2}.$$

Разлагая полученное выражение на множители и выбирая тот, у которого нули и полюса лежат в левой полуплоскости, находим

$$k_1(i2\pi f) = \sqrt{\frac{2}{N_0}}\, \frac{\alpha + i2\pi f}{\sqrt{\alpha^2 + 4\alpha P_c/N_0} + i2\pi f} =$$

$$= \sqrt{\frac{2}{N_0}}\, \frac{1 + i2\pi f/\alpha}{\sqrt{1+\Lambda} + i2\pi f/\alpha}.$$

[1] Уравнение (6.39) моделирует установившийся процесс на выходе интегрирующей RC-цепи с постоянной времени $RC = 1/\alpha$ (см. рис. 3.8), которая находится под воздействием напряжения $V(t)/\alpha$.

Далее, подставив $k^*_1(i2\pi f)$ в (6.26), находим окончательно передаточную функцию оптимального реализуемого фильтра

$$k_p = \frac{\sqrt{1+\Lambda}-1}{\sqrt{1+\Lambda}+i2\pi f/\alpha}. \qquad (6.43)$$

Такую передаточную функцию можно реализовать с помощью интегрирующей цепи с постоянной времени $RC=1/(\alpha\sqrt{1+\Lambda})$ и усилителя.

Можно показать, что в данном случае, когда сигнал и шум стационарны, фильтр Колмогорова — Винера совершенно эквивалентен фильтру Калмана (рис. 6.2*б*).

В заключение повторим, что если не накладывать на фильтр условия линейности, то можно, в принципе, построить нелинейный фильтр, обеспечивающий оценку с меньшей погрешностью. Исключение представляет случай, когда сигнал и помеха гауссовские, так как для них оптимальный фильтр всегда линеен. Вопрос о нелинейной фильтрации будет кратко затронут в § 6.10 при рассмотрении модулированных сигналов.

6.3. ОПТИМАЛЬНАЯ ОЦЕНКА НЕПРЕРЫВНЫХ ПАРАМЕТРОВ СИГНАЛА

В предыдущем параграфе рассматривалась непосредственная передача непрерывного сообщения. В большинстве современных систем связи первичный сигнал $B(t)$ не передается по каналу связи, а преобразуется в модуляторе во вторичный сигнал $S(t)$. Рассмотрим простейший случай, когда передаваемое сообщение представляет собой некоторую случайную величину, не изменяющуюся во времени. В этом случае модуляция сводится к установлению некоторого параметра λ сигнала $S(t)$ в соответствии с передаваемым сообщением, а задачей демодулятора является выявление (оценка) этого параметра с возможно большей точностью. Такая же задача возникает и при передаче сообщений, заданных на дискретных точках оси времени. В этом случае последовательно оцениваются значения параметра на интервалах времени между дискретными точками.

В простейшем случае, когда оценивается один параметр сигнала заданной формы, задача ставится следующим образом. Пусть принимаемое на интервале $(0, T)$ колебание $Z(t)$ представляет собой аддитивную смесь сигнала $S(t, \lambda)$, зависящего от одного неизвестного параметра λ, с белым шумом $N(t)$:

$$Z(t) = S(t, \lambda) + N(t), \quad 0 \leqslant t \leqslant T. \qquad (6.44)$$

Полагаем; что параметр λ имеет постоянное значение на интервале наблюдения $(0, T)$ и известна априорная плотность вероятности этого параметра $w(\lambda)$. Требуется определить оператор системы, гарантирующий получение наилучшей оценки параметра $\hat{\lambda}$, и рассчитать точность этой оценки.

Из-за шума в канале и случайного характера параметра λ точное измерение его невозможно. Можно лишь указать приближенную его оценку. Очевидно, вся информация о переданном параметре (сообщении) λ после приема колебания $Z(t)$ (6.44) содержится в апостериорном распределении $w(\lambda|z)$, которое, согласно формуле Байеса, можно представить в следующем виде:

$$w(\lambda \mid z) = w(\lambda)\, w(z \mid \lambda)/w(z). \qquad (6.45)$$

На основании анализа апостериорного распределения (6.45) принимается решение об оценке передаваемого параметра λ. При больших отношениях сигнал/шум апостериорная плотность вероятности имеет наибольший максимум в окрестности истинного значения параметра λ. Это обстоятельство указывает, что в качестве оценки целесообразно взять то значение $\hat{\lambda}$, которое обращает в максимум функцию $w(\lambda|z)$.

Во многих практических случаях априорная плотность вероятности $w(\lambda)$ оказывается неизвестной и ее полагают равномерной $w(\lambda)=\text{const}$ на некотором интервале Λ. При этом координата максимума апостериорной вероятности будет совпадать с соответствующей координатой условного распределения $w(z|\lambda)$, которое называют функцией правдоподобия. В этом случае правило максимума апостериорной вероятности переходит в правило максимального правдоподобия. Здесь оценка параметра $\hat{\lambda}_0$ определяется из условия

$$d\,w(z|\lambda)/d\lambda=0. \tag{6.46}$$

Оценка параметра, получаемая по этому критерию, называется *максимально правдоподобной.* Уравнение правдоподобия (6.46) можно записать в другом виде:

$$\partial[\ln w(z/\alpha)]/\partial\lambda=0, \tag{6.47}$$

поскольку $\ln x$ — монотонная функция своего аргумента и, следовательно, корни уравнений (6.46) и (6.47) совпадают. Оценка определяется тем корнем уравнения (6.47), который соответствует максимуму функции правдоподобия. Другим, весьма распространенным критерием оценки параметров сигнала, является оценка по минимуму среднеквадратичной ошибки. При этом критерии минимизируется по $\hat{\lambda}$ выражение

$$\overline{\varepsilon^2(\hat{\lambda})}=\int_{\Lambda}(\lambda-\hat{\lambda})^2 w(\lambda/z)\,d\lambda. \tag{6.48}$$

В этом случае оптимальная оценка $\hat{\lambda}_0$ находится из условия

$$d[\varepsilon^2(\hat{\lambda})]/d\hat{\lambda}=0.$$

После дифференцирования выражения (6.48) по $\hat{\lambda}$ с учетом, что $\int_{\Lambda} w(\lambda|z)\,d\lambda=1$, получаем $2\hat{\lambda}_0-2\int_{\Lambda}\lambda w(\lambda|z)\,d\lambda=0$, откуда

$$\hat{\lambda}_0=\int_{\Lambda}\lambda\, w(\lambda/z)\,d\lambda, \tag{6.49}$$

т. е. оптимальной оценкой параметра является математическое ожидание апостериорного распределения.

Критерий среднеквадратичной ошибки является частным случаем более общего критерия, когда минимизируется математическое ожидание некоторой функции потерь $L(\lambda-\hat{\lambda})$, т. е.

$$\overline{L(\lambda-\hat{\lambda})}=\int_{\Lambda}L(\lambda-\hat{\lambda})\,w(\lambda/z)\,d\lambda. \tag{6.50}$$

Оценка, минимизирующая эту величину, называется байесовской оценкой, а критерий (6.50), как и в дискретном случае [см. (4.13)], — критерием среднего риска. При $L(\lambda-\hat{\lambda})=(\lambda-\hat{\lambda})^2$ критерий минимума среднего риска (6.50) переходит в критерий среднеквадратической ошибки (6.48). В этом случае байесовская оценка определяется выражением (6.49).

Если $w(\lambda|z)$ симметрична относительно $\hat{\lambda}_{\text{опт}}$, что имеет место при большом отношении сигнал/шум, то критерии максимума апостериорной вероятности или максимума функции правдоподобия совпадают с критерием минимума среднеквадратичной ошибки.

Если значение параметра λ постоянно на интервале наблюдения и принятое колебание представляет собой аддитивную смесь (6.44) полезного сигнала $S(t, \lambda)$ и нормального белого шума $N(t)$ со спектральной плотностью $N_0/2$, то вектор $\mathbf{Z}$ представляющий принимаемое колебание в функциональном пространстве, является случайным гауссовским вектором, среднее значение которого равно $s(t, \lambda)$, а дисперсия совпадает с дисперсией шума. Рассуждая так же, как в § 4.2, легко показать, что

$$\ln w(\mathbf{z}|\lambda)=\frac{2}{N_0}\int_0^T z(t)\,s(t,\lambda)\,dt-\frac{1}{N_0}\int_0^T s^2(t,\lambda)\,dt. \tag{6.51}$$

Нетрудно убедиться, что для сигнала $S(t)$, зависящего от нескольких параметров $\boldsymbol{\lambda}=(\lambda_1, \lambda_2, \ldots, \lambda_n)$, функция правдоподобия

$$\ln w(\mathbf{z}/\boldsymbol{\lambda})=\frac{2}{N_0}\int_0^T z(t)\,s(t,\boldsymbol{\lambda})\,dt-\frac{1}{N_0}\int_0^T s^2(t,\boldsymbol{\lambda})\,dt. \tag{6.52}$$

Выражение для апостериорной вероятности согласно (6.51) и (6.45) будет

$$w(\boldsymbol{\lambda}/\mathbf{z})=k\,w(\boldsymbol{\lambda})\exp\left[\frac{2}{N_0}\int_0^T z(t)\,s(t;\boldsymbol{\lambda})\,dt\right]\exp\left[-\frac{1}{N_0}\int_0^T s^2(t,\boldsymbol{\lambda})\,dt\right], \tag{6.53}$$

где k — постоянный коэффициент.

Второй экспоненциальный множитель не представляет операции над $z(t)$ и может быть вынесен в виде отдельного множителя, подобного априорной вероятности. Он равен $\exp(-E|N_0)$, где E — энергия сигнала $s(t, \lambda)$. В тех случаях, когда параметр λ не энергетический, т. е. энергия сигнала не зависит от λ, этот член также можно включить в постоянную κ. При этом условии выражение (6.53) можно записать так:

$$w(\boldsymbol{\lambda}|\mathbf{z})=k_1\,w(\boldsymbol{\lambda})\exp[q(\boldsymbol{\lambda})], \tag{6.54}$$

где

$$q(\boldsymbol{\lambda})=\frac{2}{N_0}\int_0^T z(t)\,s(t,\boldsymbol{\lambda})\,dt. \tag{6.55}$$

Экспоненциальная функция является монотонной. Поэтому функция $q(\lambda)$ воспроизводит характер изменения апостериорной вероятности. Отсюда следует, что при известной априорной вероятности определение апостериорной вероятности сводится к вычислению функции $q(\lambda)$. Эта функция, с точностью до коэффициента, равна скалярному произведению пришедшего сигнала на ожидаемый вариант сигнала $s(t, \lambda)$. Ее часто называют корреляционным интегралом. Она определяет те существенные операции, которые нужно выполнить над $z(t)$, чтобы извлечь всю доступную информацию о переданном сообщении λ. Это означает, что оптимальный приемник максимального правдоподобия должен воспроизводить то сообщение λ, для которого функция $q(\lambda)$ максимальна. Определение функции $q(\lambda)$ на всем априорном интервале $\Lambda=\lambda_{макс}-\lambda_{мин}$ технически трудно реализуемо. Поэтому обычно определяется не непрерывная функция $q(\lambda)$, а совокупность ее дискретных значений $q(\lambda_1)$, $q(\lambda_2)$, ..., $q(\lambda_m)$ на интервале Λ. Тогда оптимальную схему можно реализовать в виде «многоканальной» системы (рис. 6.3), каждая ветвь которой формирует оптимальный выходной эффект (6.55) $q(\lambda_i)$ при данном фиксированном значении измеряемого параметра $\lambda_i (i=1, 2, ..., m)$. Выходные эффекты $q_i=q(\lambda_i)$ всех ветвей подводятся к решающему устройству РУ, принимающему решение $\hat{\lambda}_0$ по максимуму правдоподобия (максимуму $q(\lambda)$). Число m дискретных значений параметра λ, или число независимых ветвей приема должно быть таким, чтобы набор дискретных значений функции $q(\lambda_i)$ с достаточной для практики точностью воспроизводил функцию $q(\lambda)$ на заданном интервале Λ. Формирование выходного эффекта (корреляционного интеграла) $q(\lambda_i)$ в каждом канале можно осущест-

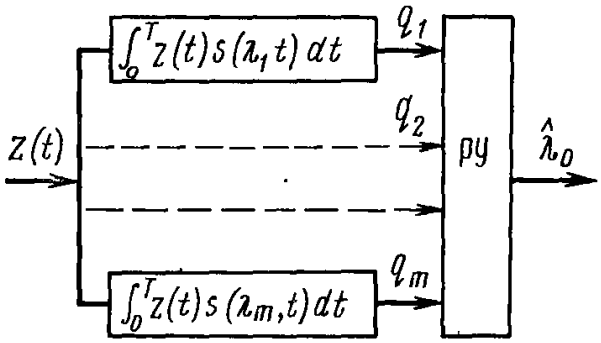


Рис. 6.3. «Многоканальная» схема оптимального приема непрерывного параметра

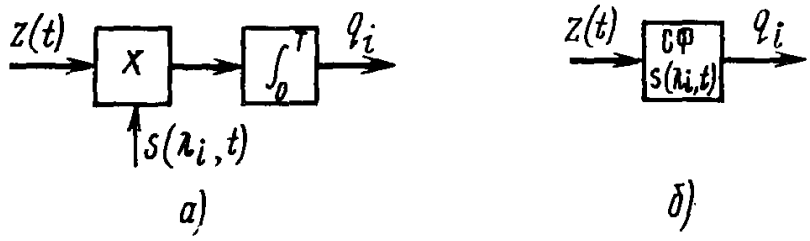


Рис. 6.4. Схемы одной ветви при корреляционном приеме (*а*) и при приеме на согласованный фильтр (*б*)

вить с помощью коррелятора (рис. 6.4*а*) или согласованного фильтра (рис. 6.4*б*).

В некоторых частных случаях возможны и более простые способы определения функции $q(\lambda)$ и нахождения оптимальной оценки $\hat{\lambda}$. Так, в радиолокации при измерении дальности цели излучают импульс $s_0(t)$ и принимают отраженный от цели импульс $s(t, \lambda) = \kappa s_0(t-\lambda)$, к которому добавляется и аддитивный шум. Здесь параметр λ — время запаздывания отраженного импульса, однозначно связанное с расстоянием до цели. Подав сумму отраженного сигнала и шума $Z(t)$ на фильтр, согласованный с $s_0(t)$, получим на выходе фильтра напряжение, совпадающее по форме с функциями взаимной корреляции $z(t)$ и $s_0(t)$ [см. (4.34)], т. е. с функцией $q(\lambda)$, сдвинутой на величину λ_0 — действительного запаздывания сигнала. Наблюдая эту функцию на экране индикатора, можно по положению ее максимума оценить λ_0, т. е. расстояние до цели.

6.4. НОРМАЛЬНЫЕ И АНОМАЛЬНЫЕ ОШИБКИ

Функцию $q(\lambda)$ (6.55) с учетом (6.44) можно представить в виде суммы двух слагаемых:

$$q(\lambda) = q_c(\lambda) + q_ш(\lambda), \tag{6.56}$$

где $q_c(\lambda) = \frac{2}{N_0}\int_0^T s(t, \lambda_0)\, s(t, \lambda)\, dt$; $q_ш(\lambda) = \frac{2}{N_0}\int_0^T N(t)\, s(t, \lambda)\, dt$; (6.57); (6.58)

$\hat{\lambda}_0$ — действительно передаваемое значение λ.

Функция $q_c(\lambda)$ представляет собой функцию автокорреляции сигнала и называется сигнальной функцией. Функция $q_ш(\lambda)$ представляет собой функцию взаимной корреляции между сигналом и шумом. Она образует как бы шумовой фон функции $q(\lambda)$ и соответственно функции $w(\lambda|z)$. Существенное различие между сигнальной и шумовой функциями состоит в том, что первая при каждом фиксированном значении λ является детерминированной, а вторая — случайной.

Рассмотрим характер этих функций с целью анализа структуры функции $q(\lambda)$, а следовательно, и структуры апостериорной вероятности $w(\lambda|z)$. Сигнальную функцию (6.57) можно записать в виде

$$q_c(\lambda) = (2E/N_0)\, R_c(\lambda), \tag{6.59}$$

где $E = \int_0^T s^2(t, \lambda)\, dt$ — энергия сигнала;

$$R_c = R_c(\lambda) = R_c(\varepsilon) = \frac{1}{E}\int_0^T s(t, \lambda_0)\, s(t, \lambda_0 + \varepsilon)\, dt \tag{6.60}$$

— нормированная корреляционная функция сигнала по передаваемому параметру; $\varepsilon = \lambda - \lambda_0$ — приращение параметра λ.

Сигнальная функция имеет максимум при $\lambda = \lambda_0$, равный $q_{c\,макс}(\lambda) = 2E/N_0$. Форма сигнальной функции $q_c(\lambda)$ определяется нормированной корреляционной функцией $R_c(\lambda)$, причем

как $R_c(\lambda)$, так и $q_c(\lambda)$ являются симметричными функциями, зависящими лишь от абсолютного значения разности $\varepsilon=\lambda-\lambda_0$. Ширину главного пика сигнальной функции (ширину области высокой корреляции сигнала) можно оценить удвоенной величиной интервала корреляции $\varepsilon_к$:

$$\varepsilon_к=\int_0^\infty |R_c(\varepsilon)|\,d\varepsilon. \tag{6.61}$$

Если шкала передаваемого сообщения нормированиа ($-1<\lambda<1$), то число независимых сигнальных пиков (число разрешимых сигналов) на интервале $\Lambda=\lambda_{макс}-\lambda_{мин}=2$ будет равно

$$m=\Lambda/2\,\varepsilon_к=1/\varepsilon_к. \tag{6.62}$$

В многоканальной схеме оптимального приема (см. рис. 6.3) это число m и определяет минимально необходимое число каналов.

Шумовая функция $q_ш(\lambda)$ (6.58), как и шум на входе $N(t)$, представляет собой случайный процесс с гауссовым распределением. Среднее значение этой функции $\overline{q_ш(\lambda)}=0$, а дисперсия

$$D\{q_ш(\lambda)\}=\overline{q_ш^2(\lambda)}=\frac{4}{N_0^2}\int_0^T\int_0^T \overline{N(t_1)\,N(t_2)}\,s(t_1,\ \lambda)\,s(t_2,\ \lambda)\,dt_1\,dt_2=\frac{2E}{N_0}.$$

Обращает на себя внимание тот факт, что максимальное значение сигнальной функции и дисперсия шумовой функции одинаковы по величине. Более подробное рассмотрение показывает, что корреляционная функция для $q_ш(\lambda)$ по форме подобна сигнальной функции $q_c(\lambda)$.

Таким образом, операция образования функции $q(\lambda)$, которая имеет место при оптимальном приеме, сопровождается «выравниванием» временных структур сигнала $q_c(\lambda)$ и шума $q_ш(\lambda)$. Однако при этом улучшаются амплитудные различия сигнала и шума, ведущие к существенному понижению шумового фона $q_ш$, на котором наблюдается сигнал q_c. Это и обеспечивает наилучшую фильтрацию сигнала из шума при оптимальном приеме.

В соответствии с изложенным выше, выражение (6.54) для апостериорной плотности вероятности $w(\lambda|\mathbf{z})$ при приеме полностью известных сигналов и $w(\lambda)=\text{const}$ можно записать в виде

$$w(\lambda|\mathbf{z})=k_2\exp\left[\frac{2E}{N_0}R_c(\lambda)\right]\exp[q_ш(\lambda)]. \tag{6.63}$$

На рис. 6.5 иллюстрируется характер зависимости апостериорной плотности вероятности $w(\lambda|\mathbf{z})$ от отношения сигнал/шум $h^2=E/N_0$. На участке вблизи истинного значения параметра $\lambda=\lambda_0$ апостериорная плотность вероятности определяется в основном функцией $R_c(\lambda)$, тогда как на остальных участках она определяется случайной функцией $q_ш(\lambda)$ и состоит из последовательности случайно возникающих выбросов, форма которых близ-

ка к $R_c(\lambda)$. Величина этих «помеховых» выбросов зависит от отношения сигнал/шум h^2 на входе приемника. При слабой входной помехе ($h^2 \gg 1$) функция $w(\lambda|z)$ имеет большой сигнальный выброс, а шумовые выбросы очень малы.

Если бы помехи в канале вовсе не было, то функция $q(\lambda)$ состояла бы только из сигнальной составляющей и ее максимум, соответствующий точному значению параметра $\lambda=\lambda_0$, можно было бы определить без ошибки. При слабых помехах функция $q(\lambda)$ имеет также лишь один значительный выброс, определяемый сигнальной составляющей $q_c(\lambda)$. Однако на него накладывается и шумовая составляющая $q_ш(\lambda)$, которая несколько изменяет форму выброса и может сместить его максимум. В результате параметр λ определяется с ошибкой, которая с вероятностью, близкой к единице, не превышает по абсолютной величине интервала корреляции $\varepsilon_к$ (6.61). Ошибки, не выходящие из области сигнального пика ($|\varepsilon|<\varepsilon_к$), будем называть *нормальными*[1]. Увеличение мощности помехи на входе смещает и расширяет сигнальный пик. Кроме этого, растут шумовые выбросы. В результате наибольший из них может быть принят за сигнальный, что приведет к появлению ненормально больших ошибок. Ошибки, превышающие по абсолютной величине значения интервала корреляции $\varepsilon_к$ (6.61), называются *аномальными*. Появление аномальных ошибок приводит к значительному уменьшению верности и является причиной возникновения порога помехоустойчивости в широкополосных системах модуляции (см. § 6.7). Таким образом, в области сильных помех помехоустойчивость передачи непрерывных сообщений определяется вероятностью аномальных ошибок $p_{ап}=P(|\varepsilon|>\varepsilon_к)$. При передаче непрерывного сообщения вероятность аномальной ошибки определяет *достоверность* оценки параметра, а нормальная ошибка определяет *точность* (качество) этой оценки.

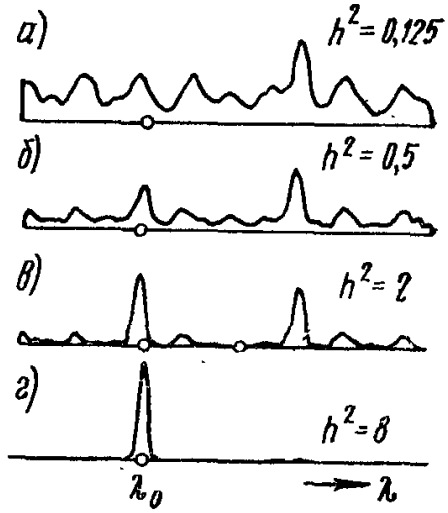


Рис. 6.5. Характер зависимости апостериорных распределений для различных значений отношения сигнал/шум

В соответствии с (6.62) передачу непрерывного параметра можно заменить дискретной передачей m взаимно ортогональных

[1] Здесь термин «нормальная ошибка» относится не к распределению вероятностей, а связан с тем фактом, что в нормальных условиях связи такие ошибки всегда существуют. Рассматриваемые ниже аномальные ошибки, наоборот, в нормальных условиях появляются крайне редко и оказываются существенными только в ненормальных условиях, когда верность связи падает ниже допустимого уровня. Впрочем, очень часто апостериорная вероятность $w(\lambda|z)$ вблизи сигнального пика описывается гауссовской кривой, так что нормальные ошибки и с точки зрения распределения вероятностей оказываются нормальными.

сигналов. Тогда вероятность аномальной ошибки при передаче непрерывных сообщений приближенно можно определить как вероятность ошибки в эквивалентной системе дискретной передачи m ортогональных сигналов. Поэтому для расчета $p_{ан}$ остается только воспользоваться соответствующими формулами, полученными в гл. 4. Более точные расчеты $p_{ан}$ для различных систем передачи непрерывных сообщений сопряжены с определенными математическими трудностями.

Определим среднюю квадратичную погрешность максимально правдоподобной оценки $\hat{\lambda}$ при малой интенсивности шума. Из (6.54) видно, что $\hat{\lambda}$ — это то значение λ, при котором $q(\lambda)$ принимает наибольшее значение. Здесь полезно привести геометрическую интерпретацию. В пространстве сигналов каждой реализации сигнала $s(t, \lambda)$ при различных λ соответствует точка. Если s зависит непрерывно от λ (что имеет место во всех аналоговых системах связи), то все эти точки образуют некоторую кривую. Принятый сигнал $z(t)$ является также точкой в пространстве сигналов, как правило, не лежащей на кривой $s(t, \lambda)$ (рис. 6.6). Максимально правдоподобная оценка $\hat{\lambda}$ соответствует тому сигналу $s(t, \hat{\lambda})$, который изображается на сигнальной кривой точкой, ближайшей к точке $z(t)$. Обозначим $\Delta s = s(t, \hat{\lambda}) - s(t, \lambda_0)$, где λ_0 — действительно переданный параметр.

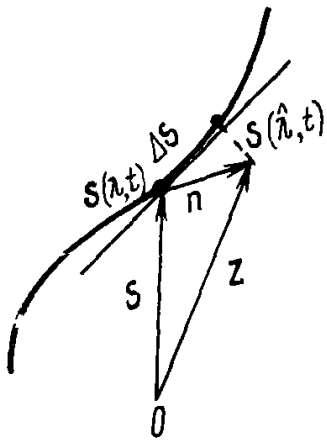


Рис. 6.6. Геометрическое представление сигнала и шума

При малой помехе и, следовательно, малом отклонении Δs можно считать отрезок между $s(t, \lambda_0)$ и $s(t, \hat{\lambda})$ прямой. Тогда Δs представляет проекцию вектора помехи $\mathbf{n}$ на эту прямую. Докажем, что $\overline{\Delta s^2} = N_0/(2T)$, где N_0 — спектральная плотность помехи; T — длительность сигнала. Для этого разложим случайный вектор помехи (квазибелого шума) в $2FT$-мерном пространстве на $2FT$ ортогональных составляющих, одной из которых является Δs. Очевидно, что все эти составляющие являются нормальными центрированными случайными величинами с одинаковыми дисперсиями σ^2. Сумма этих дисперсий равна мощности квазибелого шума в полосе частот F, т. е. $2FT\sigma^2 = FN_0$, откуда

$$\sigma^2 = \overline{\Delta s^2} = N_0/(2T). \tag{6.64}$$

С другой стороны, при малых Δs можно полагать

$$\Delta s \approx \Delta\lambda \frac{\partial s(t, \lambda)}{\partial \lambda}, \tag{6.65}$$

где $\Delta\lambda = \hat{\lambda} - \lambda_0$ — погрешность оценки параметра. Отсюда

$$\overline{\Delta\lambda^2} \approx \frac{\overline{\Delta s^2}}{\overline{[\partial s(t, \lambda)/\partial\lambda]^2}} = \frac{N_0}{2T\,\overline{[\partial s(t, \lambda)/\partial\lambda]^2}}. \tag{6.66}$$

Здесь усреднение квадрата частной производной производится и по времени и по ансамблю λ.

Из полученного результата видно, что погрешность в оценке параметра зависит от вида функции $s(t, \lambda)$, т. е. от сущности параметра λ. Чем больше $\overline{[\partial s(t, \lambda)/\partial\lambda]^2}$ при фиксированной мощности сигнала, тем более помехоустойчива система.

Рассмотрим некоторые виды модуляции при передаче отдельных значений непрерывного параметра λ. Их можно подразделить на линейные и нелинейные, в зависимости от вида функции $s(t, \lambda)$.

Любой линейный вид модуляции можно представить в форме $S(t, \lambda)=\varphi(t)+\lambda\psi(t)$, где $\varphi(t)$ и $\psi(t)$ — некоторые детерминированные функции времени. В этом случае $\partial s/\partial\lambda=\psi(t)$ и

$$\overline{\Delta\lambda^2}=\frac{N_0}{2T\overline{\psi^2(t)}}=\frac{N_0}{2TP_\psi}=\frac{N_0}{2E_\psi}, \tag{6.67}$$

где $P_\Psi=\psi^2(t)$ — мощность составляющей сигнала $\psi(t)$; $E_\Psi=P_\Psi T$ — ее энергия.

Пусть, например, параметр λ меняется от -1 до $+1$, а $\varphi(t)=\psi(t)$, т. е. $s(t, \lambda)=\psi(t)[1+\lambda]$ (амплитудная модуляция). При этом $E_\Psi=\frac{1}{4}E_s$, где E_s — энергия сигнала при передаче максимального параметра, и $\overline{\Delta\lambda^2}=2N_0/E_s$.

Более высокую помехоустойчивость можно получить, положив $\varphi(t)=0$. Тогда $s(t, \lambda)=\lambda\psi(t)$ (балансная модуляция), $E_\Psi=E_s$ и $\overline{\Delta\lambda^2}=N_0/2E_s$.

Легко видеть, что более высокой помехоустойчивости при линейной модуляции получить нельзя. Величина погрешности оценки не зависит от вида функции $\psi(t)$. Обычно используют гармонический сигнал $\psi(t)=u_0\cos(\omega t+\theta)$.

Отметим также, что при линейной модуляции сигнальная линия в пространстве сигналов прямая, а равенство (6.65) справедливо для любых Δs, а не только для малых. Поэтому полученные результаты для линейной модуляции верны не только при слабых помехах, но и при сильных. Другими словами, при линейной модуляции и оптимальном приеме аномальных ошибок не бывает.

При нелинейной модуляции аномальные ошибки при больших отношениях помехи к сигналу возникают неизбежно, однако при слабых помехах можно получить значительно более высокую помехоустойчивость, чем у линейной модуляции. Приведем два примера.

а) Частотная модуляция. Здесь $s(t, \lambda)=u_0\cos[(\omega_0+\Delta\omega\lambda)t]$, $-T/2<t<T/2$, где $\Delta\omega$ — максимальное абсолютное отклонение (девиация) частоты. При этом

$$\overline{(\partial s/\partial\lambda)^2}=\overline{[-\Delta\omega t u_0\sin(\omega_0+\Delta\omega\lambda)t]^2}=$$

$$=\frac{1}{2}\Delta\omega^2 u_0^2(T^2/12)=\Delta\omega^2 P_c(T^2/12),$$

где $P_c = u_0^2/2$ — мощность сигнала. Отсюда

$$\overline{\Delta\lambda^2} = \frac{12 N_0}{2T^3 \Delta\omega^2 P_c} = \frac{6N_0}{E_c \Delta\omega^2 T^2}. \qquad (6.68)$$

Выбрав $\Delta\omega T \gg \sqrt{12}$, можно получить помехоустойчивость значительно выше, чем при линейной модуляции. Заметим, что это достигается за счет использования более широкой полосы частот.

б) Время импульсная модуляция. Здесь $s(t, \lambda) = s_0(t+\lambda\tau)$, где $s_0(t)$ — любой детерминированный импульсный сигнал; τ — максимальное отклонение момента прихода сигнала. В этом случае $\overline{(\partial s/\partial\lambda)^2} = \tau^2 \overline{(\partial s_0/dt)^2}$ и

$$\overline{\Delta\lambda^2} = \frac{N_0}{2\tau^2 T \overline{(ds_0/dt)^2}}. \qquad (6.69)$$

Помехоустойчивость ВИМ при слабых помехах определяется, таким образом, энергией производной импульса ds_0/dt. Но ее можно сделать сколь угодно большой, не увеличивая энергии самого импульсного сигнала. Для этого достаточно, например, применить импульс с фронтами, перпендикулярными оси времени. Казалось бы, что таким образом можно полностью избавиться от действия помех. К сожалению, это не так. Чем круче фронты импульса, тем меньше тот уровень помех, при котором возникают аномальные ошибки. Для точно прямоугольного импульса любая помеха является «сильной» и формула (6.66), полученная для слабых помех, неприменима.

6.5. ОПТИМАЛЬНЫЙ ПРИЕМ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Определим условия оптимального приема непрерывных сообщений (функций времени). Пусть сообщение представляет собой некоторый процесс (первичный сигнал) $B(t)$. Он может непрерывно изменяться со временем и принимать любую форму. Такие сообщения встречаются, например, в телефонии, телевидении, телеметрии.

Для простоты анализа будем считать, что функция $B(t)$ принимает значения в пределах от $+1$ до -1. Будем также полагать, что реализации сообщения $b(t)$ принадлежат ν-мерному евклидову пространству и, следовательно, могут быть разложены по ортонормированному базису $\{\varphi_k(t)\}$:

$$b(t) = \sum_{k=1}^{\nu} \lambda_k \varphi_k(t), \qquad (6.70)$$

где λ_k — случайные параметры, определяющие передаваемую информацию. Обычно пространство R_ν, к которому принадлежат сообщения, — это пространство финитных функций длительностью T, представимых тригонометрической суммой (см. с. 62).

Тогда $\varphi_k(t)$ — отрезки гармонических функций:

$$b(t)=\sum_{i=1}^{\nu/2}\left(\lambda_{2i-1}\sqrt{2}\sin\frac{2\pi i}{T}t+\lambda_{2i}\sqrt{2}\cos\frac{2\pi i}{T}t\right),$$

$$-T/2<t<T/2;\quad \nu=2F_cT. \tag{6.71}$$

Для передачи по каналу колебание $b(t)$ преобразуется в сигнал $s(t, b)$. Поскольку колебание (6.70) определяется ν параметрами λ_k, то и сигнал будет зависеть от этих параметров. Принятый сигнал вследствие наложения помехи $N(t)$ равен

$$Z(t)=S(t, B)+N(t)=S(t, \lambda)+N(t).$$

Таким образом, задача оптимального приема непрерывного сообщения $b(t)$ сводится к задаче совместного оптимального приема совокупности многих параметров $\lambda=(\lambda_1, \lambda_2, \ldots, \lambda_\nu)$. Эта задача является обобщением рассмотренной в § 6.4 задачи оптимальной оценки одного параметра.

Итак, по реализации $z(t)$ необходимо восстановить переданное сообщение $b(t)$ с возможно большей точностью, хотя бы при слабых помехах. Как мы уже установили, для этого необходимо на основе анализа принятого колебания $z(t)$ найти максимум в пространстве R_ν апостериорного распределения:

$$w(\mathbf{b}\,|\,\mathbf{z})=k\,w(\mathbf{b})\exp\left\{-\frac{1}{N_0}\int_0^T[z(t)-s(t,\ \mathbf{b})]^2\,dt\right\} \tag{6.72}$$

или при $w(\mathbf{b})=\text{const}$ — минимум по $b(t)$ интеграла

$$d^2=\int_0^T[z(t)-s(t,\ \mathbf{b})]^2\,dt. \tag{6.73}$$

Таким образом, оптимальный приемник должен воспроизводить сообщение $\hat{b}(t)$, при котором среднеквадратическое отклонение d имеет минимальное значение. При отсутствии помех такой приемник воспроизводит сообщение без искажений (без ошибок): $z(t)=s(t, b)$, $\hat{b}(t)=\mathbf{b}(t)$ и $\overline{\varepsilon^2(t)}=0$, а при наличии помех ошибка будет минимальной.

Из условия (6.73) следует, что операции оптимальной фильтрации и детектирования дают достаточное решение задачи об извлечении максимальной информации из принятого сигнала $z(t)$ относительно переданного сообщения $b(t)$. Структурная схема приемника с оптимальным фильтром приведена на рис. 6.7.

Теория линейной фильтрации Колмогорова — Винера для нормальных стационарных процессов была рассмотрена в § 6.2. Практическая реализация оптимальных фильтров сопряжена с большими трудностями. А если учесть, что реальные модулированные сигналы не являются стационарными нормальными процессами, то задача построения оптимального фильтра на базе линейной теории становится практически неразрешимой. Поэтому представляет интерес рассмотрение других путей построения оптимального приемника непрерывных сообщений.

Запишем выражение (6.72) в другом виде, подобном выражению (6.54):

$$w(\mathbf{b} \mid \mathbf{z}) = \kappa\, w(\mathbf{b}) \exp[q(\mathbf{b})], \tag{6.74}$$

где

$$q(\mathbf{b}) = \frac{2}{N_0} \int_0^T z(t)\, s(t, \mathbf{b})\, dt. \tag{6.75}$$

Отсюда следует, что при известной априорной вероятности определение апостериорной вероятности сводится к вычислению функции $q(\mathbf{b})$, т. е. скалярного произведения принятого колебания $z(t)$ на переданные (ожидаемые) сигналы $s(t, \mathbf{b})$. Такая обработка может быть реализована с помощью многоканального устройства, аналогичного устройству, показанному на рис. 6.3. Однако схема этого устройства теперь существенно усложнится вследствие того, что передаваемое сообщение $b(t)$ определяется не одним, а большим числом параметров.

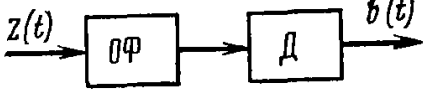


Рис. 6.7. Структурная схема приемника непрерывных сигналов

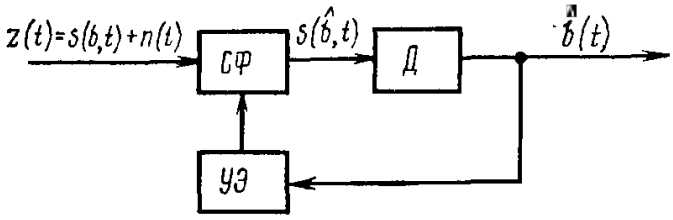


Рис. 6.8. Структурная схема приемника со следящим фильтром

Во многих случаях целесообразно применение более простых, следящих устройств. Рассмотрим принципиальную возможность построения таких устройств. Подробное и более строгое обоснование на основе теории нелинейной фильтрации приводится в § 6.10.

При передаче непрерывных сообщений сигнал $s(t, \mathbf{b})$ не является полностью известным. Однако обычно имеется некоторая априорная информация об этом сигнале. Известны, например, несущая частота, вид модуляции, ширина спектра сигнала и т. п. Часть информации можно получить в результате наблюдения над принятой реализацией сигнала $z(t)$ за предшествующий промежуток времени. В результате имеется возможность определить оценку сигнала $s(t, \hat{\mathbf{b}})$ и вычислить функцию $q(\hat{\mathbf{b}})$ для этой оценки:

$$q(\hat{\mathbf{b}}) = \frac{2}{N_0} \int_0^T z(t)\, s(t, \hat{\mathbf{b}})\, dt. \tag{6.76}$$

Функцию $q(\hat{b})$ можно вычислить с помощью фильтра с переменными параметрами (рис. 6.8) или схемы следящего коррелятора (рис. 6.9). Каждая из этих схем имеет основной информационный канал, на выходе которого получается оценочное значение $\hat{\mathbf{b}}(t)$ передаваемого сообщения, и канал обратной связи, с помощью которого в схеме рис. 6.9 формируется опорный сигнал $s(t, \hat{\mathbf{b}})$, а в схеме рис. 6.8 производится изменение параметров фильтра. В схеме рис. 6.8 с помощью управляющего элемента УЭ производятся изменения параметров фильтра СФ так, чтобы он был согласован с непрерывно изменяющимся ожидаемым сигналом $s(t, \hat{\mathbf{b}})$. В схеме рис. 6.9 с помощью УЭ изменяется модулируемый параметр несущего колебания, формируемого генератором Г. При частотной модуляции, например, этим

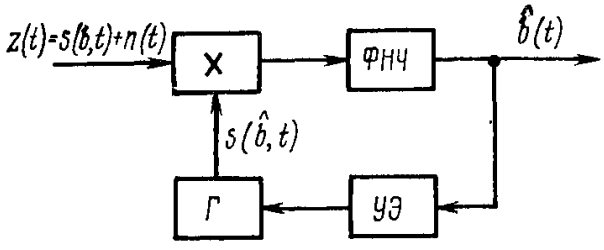


Рис. 6.9. Структурная схема следящего корреляционного приемника

параметром будет частота, при временной импульсной модуляции — сдвиг импульсов по времени и т. п. Фильтр нижних частот ФНЧ в этой схеме выполняет роль интегратора на интервале наблюдения T, который связан с максимальной частотой F_c в спектре передаваемого сообщения соотношением $T=1/2F_c$.

Рассмотренные схемы являются квазиоптимальными, поскольку получаемая оценка $s(t, b)$ не является наилучшей возможной. При различных видах модуляции принцип следящего приема остается одним и тем же. Вид модуляции определяет параметр, за которым должно осуществляться слежение. Иначе говоря, оптимальный приемник должен с наименьшей ошибкой следить за передаваемым случайным колебанием $b(t)$. Схемы следящего приема позволяют практически реализовать помехоустойчивость, близкую к потенциальной.

Перейдем к определению помехоустойчивости систем связи при оптимальном приеме. Заметим, что эту потенциальную помехоустойчивость можно вычислить, не уточняя структуры оптимального демодулятора. Для этого достаточно знать, что он выдает решение $\tilde{b}(t)$, соответствующее максимуму (6.74).

Прежде чем приступить к выводу формул, определяющих потенциальную помехоустойчивость, напомним основные принципы классификации видов модуляции при передаче непрерывных сообщений. В общем случае модуляция заключается в том, что множество сообщений (первичных сигналов) $B(t)$ преобразуется (отображается) в множество вторичных сигналов $S(t) = S[t, B(t)]$. Этой записью подчеркивается, что значение сигнала S в некоторый момент t_0 определяется в общем случае всем поведением сообщения $B(t)$ на всей оси времени.

В частном случае, если сигнал $S(t)$ в любой момент t_0 зависит не от всего хода сигнала $B(t)$, а только от его значения в момент t_0, то система модуляции называется *прямой*. К прямым относится подавляющее большинство применяемых методов модуляции. Остальные системы модуляции, в которых $S(t)$ зависит от общего поведения сигнала $B(t)$, называются непрямыми. Среди них особый интерес представляют *интегральные* системы, в которых $B(t_0)$ входит в выражение $S(t)$ под интегралом.

Система модуляции называется *линейной*, если $S(t)$ можно получить из $B(t)$ с помощью линейных операций. Линейные системы могут быть прямыми (например, амплитудная — АМ) и непрямыми (например, однополосная — ОМ).

Геометрически модуляцию можно рассматривать как отображение пространства сообщений B в пространство сигналов S, а демодуляцию — как обратное отображение. При демодуляции помеха $N(t)$ на входе приемника отображается в погрешность оценки сообщения (шум воспроизведения или шум на выходе приемника) $\varepsilon(t)$.

Рассматривая прием непрерывного сообщения на фоне белого гауссовского шума как оценку совокупности параметров λ_k в представлении (6.70) и повторяя рассуждения предыдущего параграфа, можно прийти к следующим результатам. При достаточно слабых помехах погрешность $\varepsilon(t)=\tilde{B}(t)-B(t)$, или шум воспроизведения на выходе приемника представляет собой гауссовский процесс, спектральная плотность которого определяется

так: по формуле (6.66) можно вычислить погрешность $\overline{\Delta\lambda^2_k}$ оценки каждого параметра λ_k в выражении (6.71), а мощность шума на выходе приемника в полосе $\Delta F=1/T$ вокруг частоты $f_k=k\omega_0/2\pi$ будет равна $\overline{\Delta\lambda^2_{k1}+\Delta\lambda^2_{k2}}$, где λ_{k1} и λ_{k2} — соответственно коэффициенты при косинусной и синусной составляющих на частоте $k\omega_0$ в (6.71). Тогда с учетом (6.66) спектральная плотность шума на выходе при $f_k=k\omega_0/2\pi$ будет

$$G(kf)=T\left(\Delta\lambda^2_{k1}+\Delta\lambda^2_{k2}\right)=\frac{N_0}{2}\,\overline{[1/(\partial S/\partial\lambda_{k1})^2+1/(\partial S/\partial\lambda_{k2})^2]}.$$

В большинстве практических случаев для аналоговых видов модуляции $\overline{(\partial S/\partial\lambda_{k1})^2}=\overline{(\partial S/\partial\lambda_{k2})^2}$, и тогда $G(kf)=N_0/\overline{(\partial S/\partial\lambda_k)^2}$.

Для прямых систем модуляции, в которых сообщение $b(t)$ входит непосредственно в выражение сигнала $S(b, t)$ можно записать

$$\frac{\partial S}{\partial\lambda_k}=\frac{\partial S}{\partial B}\frac{\partial B}{\partial\lambda_k},\quad\text{но}\quad\frac{\partial B}{\partial\lambda_k}=\sqrt{2}\cos k\omega_0 t$$

и $$\frac{\partial S}{\partial\lambda_k}=\sqrt{2}\frac{\partial S}{\partial B}\cos k\omega_0 t.$$

Тогда

$$\overline{\left(\frac{\partial S}{\partial\lambda_k}\right)^2}=2\cos^2 k\omega_0 t\,\overline{\left(\frac{\partial S}{\partial B}\right)^2}=\overline{\left(\frac{\partial S}{\partial B}\right)^2},$$

а спектральная плотность мощности шума на выходе приемника

$$\hat{G}_{\text{ш}}(f)=N_0\Big/\overline{\left(\frac{\partial S}{\partial B}\right)^2}. \tag{6.77}$$

Таким образом, при прямых системах модуляции шум на выходе приемника квазибелый, т. е. имеет равномерный спектр в полосе частот F_c.

В случае интегральных систем сообщение $B(t)$ входит в выражения сигнала под знаком интеграла: $S(B, t)=S[\int_0^t B(\tau)d\tau, t]=S(\psi, t)$, где $\psi(t)=\int_0^t B(\tau)d\tau$. Так как $\hat{B}(t)=B(t)+\varepsilon(t)$, то $\hat{\psi}(t)=\int_0^T B(t)dt+\int_0^t\varepsilon(t)dt=\psi(t)+\xi(t)$. Следовательно, энергетический спектр шума на выходе приемника для таких систем можно определить как энергетический спектр производной $\xi'(t)$. На основании известной теоремы о спектре производной $G_{\xi'}(f)=(2\pi f)^2 G_\xi(f)$, где $G_\xi(f)$ определяется по формуле (6.77), если в последней вместо $\partial S/\partial B$ подставить $\partial S/\partial\psi$. Таким образом, для интегральных систем энергетический спектр шума на выходе приемника будет

$$G_{\text{ш}}(f)=(2\pi f)^2 N_0/\overline{(\partial S/\partial\psi)^2}, \tag{6.78}$$

т. е. энергетический спектр помехи на выходе приемника в интегральных системах — параболический (пропорционален квадрату частоты).

Все эти результаты получены, как и в предыдущем параграфе, для слабых помех, когда можно считать $\Delta S = \Delta\lambda \frac{dS}{d\lambda}$. Они характеризуют нормальные ошибки.

Мощность шума на выходе приемника в полосе частот от нуля до F_c, очевидно, равна $P_в = \int_0^{F_c} G_ш(f)df$.

С другой стороны, мощность P_B сообщения на выходе приемника, равную $\overline{B^2(t)}$, можно выразить через пик-фактор сообщения $П = |B(t)|_{макс}/\sqrt{\overline{B^2(t)}}$. Учитывая принятое нормирование сообщения $|B(t)|_{макс} = 1$, получим

$$B^2(t) = P_B = 1/П^2. \qquad (6.79)$$

Заметим, что для гармонического сигнала $П = \sqrt{2}$, а для телефонного сообщения $П \approx 3$.

В качестве меры верности передачи непрерывного сообщения можно принять отношение его мощности к мощности шума на выходе приемника:

$$\rho_{вых} = \frac{P_B}{P_в} = \frac{1}{П^2 \int_0^{F_c} \hat{G}_ш(f)\,df}.$$

Обозначим $\rho_{вх} = P_c/P_ш = P_c/N_0F$ — отношение мощности сигнала к мощности шума (в полосе частот сигнала F) на входе приемника.

Отношение

$$g = \frac{\rho_{вых}}{\rho_{вх}} = \frac{N_0 F}{П^2 P_c \int_0^{F_c} \hat{G}_ш(f)\,df} \qquad (6.80)$$

называется *выигрышем* системы модуляции и может служить мерой ее помехоустойчивости. При $g > 1$ отношение сигнал/помеха в процессе демодуляции возрастает. Заметим, впрочем, что в некоторых случаях $g < 1$, так что вместо выигрыша система модуляции может дать проигрыш.

При сравнительной оценке различных систем связи неудобно пользоваться величиной выигрыша. Дело в том, что в различных системах сигналы могут иметь различную ширину спектра. Если помеха в канале имеет постоянную спектральную плотность, то мощность помехи на входе приемника будет выше для системы с более широкополосным сигналом. Если две системы связи в одном и том же канале при одинаковой мощности сигнала обеспечивают одинаковое качество воспроизведения сообщения (одинаковое $\rho_{вых}$), то их следует считать равноценными по помехоус-

тойчивости. В то же время, если сравнивать их по величине выигрыша g, то придется отдать предпочтение системе с более широкополосным сигналом, так как у нее $\rho_{вх}$ меньше. При переходе от узкополосной системы к широкополосной наряду с выигрышем g имеет место и проигрыш, поскольку мощность шума на входе демодулятора возрастает пропорционально расширению спектра сигнала.

Это противоречие можно устранить, если сравнивать отношения мощностей сигнала не к мощностям помехи, а к их средним спектральным плотностям, т. е. определять *реальный* или *«обобщенный выигрыш системы»*.

$$g' = \frac{\widetilde{\rho}_{вых}}{\widetilde{\rho}_{вх}} = \frac{\rho_{вых}}{\alpha \rho_{вх}} = \frac{g}{\alpha} = \frac{N_0 F_c}{\Pi^2 P_c \int_0^{F_c} \hat{G}_{ш}(f)\, df}, \tag{6.81}$$

где $\widetilde{\rho}_{вых} = \dfrac{P_B}{P_\varepsilon / F_c} = \rho_{вых} F_c$, $\widetilde{\rho}_{вх} = \dfrac{P_c}{P_ш / F} = \dfrac{P_c}{N_0} = \rho_{вх} F$, $\alpha = F/F_c$.

Заметим, что в системе непосредственной передачи, в которой сигнал пропорционален передаваемому сообщению $S(t) = kB(t)$, а $F = F_c$ и $\rho_{вых} = \rho_{вх}$, выигрыш (6.80) и обобщенный выигрыш (6.81) одинаковы и равны единице: $g = g' = 1$. Это означает, что введенный критерий выигрыша сводится к сравнению данной системы передачи с системой непосредственной передачи.

Используя теорему Шеннона (см. § 6.2), можно найти максимальные возможные значения выигрыша и обобщенного выигрыша при заданных параметрах системы связи. Рассмотрим этот вопрос для наиболее простого случая, когда непрерывное сообщение представляет гауссовский процесс с равномерным спектром в полосе частот F_c (квазибелый шум), а в канале существует аддитивная помеха в виде белого шума с односторонней спектральной плотностью N_0. Согласно теореме Шеннона передача сообщения с заданным значением $P_B/P_\varepsilon = \rho_0$ возможна в случае, когда $H'_\varepsilon(B) < C$. Здесь $H'_\varepsilon(B)$ — эпсилон-производительность источника, которая в данном случае согласно (6.9) равна $F_c \log \rho_0$, а C — пропускная способность гауссовского канала, равная согласно (3.64)

$$F \log\left(1 + \frac{P_c}{N_0 F}\right) = F \log(1 + \rho_{вх}).$$

В гипотетической *идеальной системе связи*, в которой полностью используется пропускная способность канала и $\rho_{вых} = \rho_0$,

$$F_c \log \rho_{вых} = F \log(1 + \rho_{вх}). \tag{6.82}$$

В реальных системах связи обычно удается лишь частично использовать пропускную способность канала. Назовем *эффективностью η системы модуляции* отношение эпсилон-производительности источника к минимальной пропускной способности канала,

при которой обеспечивается заданная верность, т. е. $\rho_{вых}=\rho_0$. Для такой реальной системы вместо (6.82) имеем

$$F_c \log \rho_{вых} = \eta F \log (1+\rho_{вх}). \tag{6.83}$$

Из этих выражений видно, что при $\eta F > F_c$ можно обеспечить высокую верность (большое значение $\rho_{вых}$) при относительно малых $\rho_{вх}$, т. е. получить большой выигрыш g.

Таким образом, выигрыш достигается в результате обмена ширины спектра на динамический диапазон, о чем говорилось в § 1.2. Большой выигрыш можно получить только при большом отношении $\alpha=F/F_c$. Заметим, что большой выигрыш может иметь место и при малой эффективности η и наоборот. Следовательно, при оценке различных систем связи необходимо учитывать, по крайней мере, два показателя: эффективность и помехоустойчивость. Совокупность этих двух показателей составляет достаточно полную характеристику системы.

Наилучшей системой считается такая, которая обеспечивает наибольшую помехоустойчивость при заданной эффективности либо наибольшую эффективность при заданной помехоустойчивости.

Система, в которой пропускная способность канала используется полностью ($\eta=1$), называется идеальной (ИС). Для такой системы из (6.82) следует

$$\rho_{вых} = (1+\rho_{вх})^{\alpha}. \tag{6.84}$$

Отсюда при $\rho_{вх} \gg 1$ получаем

$$g \approx \rho_{вх}^{\alpha-1}; \quad g' \approx \frac{1}{\alpha}\rho_{вх}^{\alpha-1}. \tag{6.85}$$

Таким образом, в идеальной системе выигрыш g возрастает с увеличением α по экспоненциальному закону. Никакая реальная система не может обеспечить при заданном α более высокую помехоустойчивость, чем идеальная система.

6.6. ПОМЕХОУСТОЙЧИВОСТЬ СИСТЕМ АНАЛОГОВОЙ ПЕРЕДАЧИ ПРИ СЛАБЫХ ПОМЕХАХ

Для передачи непрерывных сообщений по каналу связи применяются различные виды модуляции. При синусоидальном переносчике на практике используются амплитудная модуляция (АМ), балансная модуляция (БМ), однополосная модуляция (ОМ), фазовая модуляция (ФМ) и частотная модуляция (ЧМ). Определим потенциальную и реальную помехоустойчивость этих систем.

Амплитудная модуляция формирует сигнал вида $S(B, t)=u_0[1+mB(t)]\cos\omega_0 t$. Тогда

$$\frac{\partial S(B,\ t)}{\partial B} = m u_0 \cos \omega_0 t; \quad \overline{\left(\frac{\partial S}{\partial B}\right)^2} = \frac{1}{T}\int_0^T \overline{\left(\frac{\partial S}{\partial B}\right)^2} dt =$$

$$= \frac{1}{T}\int_0^T m^2 u_0^2 \cos^2 \omega t\, dt = \frac{1}{2T} m^2 u_0^2 \int_0^T [1 + \cos 2\omega_0 t]\, dt = \frac{1}{2} m^2 u_0^2.$$

Согласно (6.77) энергетический спектр шума на выходе приемника АМ

$$G_{\text{ш}}(f) = \frac{2N_0}{m^2 u_0^2}. \tag{6.86}$$

Мощность сигнала на входе приемника [1]

$$P_{\text{с}} = \overline{S^2(B,\ t)} = \frac{u_0^2}{T}\int_0^T \overline{[1 + m B(t)]^2} \cos^2 \omega_0 t\, dt =$$

$$= \frac{u_0^2}{2T}\int_0^T [1 + 2m\overline{B(t)} + m^2 \overline{B^2(t)}]\,[1 + \cos 2\omega_0 t]\, dt.$$

Поскольку сообщение — центрированный процесс, $\overline{B(t)} = 0$, а вследствие принятой нормировки $B^2(t) = [B_{\text{макс}}]^2/\Pi^2 = \frac{1}{\Pi^2}$,

$$P_{\text{с}} = \frac{u_0^2}{2}\left(1 + \frac{m^2}{\Pi^2}\right).$$

Здесь и далее полагаем, что $\omega_0 T \gg 1$ и поэтому значением $\int_0^T \cos 2\omega_0 t\, dt = \frac{\sin 2\omega_0 T}{2\omega_0} < \frac{1}{2\omega_0}$ можно пренебречь по сравнению с T.

На основании (6.80) (6.81) с учетом, что при АМ $F = 2F_{\text{с}}$, получаем следующие выражения для выигрыша и обобщенного выигрыша при оптимальном приеме:

$$g_{\text{АМ}} = \frac{\alpha m^2}{m^2 + \Pi^2} = \frac{2m^2}{m^2 + \Pi^2}; \qquad g'_{\text{АМ}} = \frac{m^2}{m^2 + \Pi^2}. \tag{6.87}$$

Предельное значение выигрыша при АМ равно 1. Оно достигается тогда, когда $m = 1$ и $\Pi = 1$. Практически всегда $m < 1$, а $\Pi > 1$, и поэтому $g_{\text{АМ}} < 1$ и $g'_{\text{АМ}} < 0{,}5$, т. е. система АМ дает не выигрыш, а проигрыш. Так, при передаче речи $\Pi \approx 3$, и тогда при $m = 0{,}9$ $g_{\text{АМ}} = 0{,}165$ и $g'_{\text{АМ}} = 0{,}083$. Малые значения выигрыша при АМ обусловлены тем, что лишь небольшая часть мощности сигнала заключена в боковых полосах, несущих полезную информацию. Следовательно, устранение несущей в сигнале АМ может

[1] Модулированный сигнал $S(t)$ обычно является нестационарным процессом, и поэтому для нахождения его мощности необходимо производить усреднение квадрата сигнала по ансамблю и по времени.

привести к увеличению выигрыша, что и имеет место при балансной и однополосной модуляции.

При балансной модуляции сигнал формируется путем простого перемножения первичного сигнала $B(t)$ и колебания переносчика $u_0 \cos \omega_0 t$:

$$S(B, t) = u_0 B(t) \cos \omega_0 t.$$

Средняя мощность такого сигнала $P_c = \frac{1}{2} \overline{u^2_0 B^2(t)} = \frac{1}{2} \frac{u^2_0}{\Pi^2}$, а $\overline{\left(\frac{\partial S}{\partial B}\right)^2} = \frac{1}{2} u^2_0$. С учетом этих соотношений имеем (при $\alpha = 2$), $g_{БМ} = 2$, $g'_{БМ} = 1$. Отсюда, выигрыш при БМ не зависит от пик-фактора сообщения.

Однополосная модуляция, как известно, представляет собой просто перенос спектра первичного сигнала вверх на частоту ω_0. Это линейная операция, при которой ширина спектра не изменяется и не изменяются также соотношения между мощностями составляющих. Поэтому при демодуляции входная помеха преобразуется в выходную таким же образом, как и сигнал. Из этих соображений можно, не прибегая к громоздким вычислениям, записать: $g_{ОМ} = g'_{ОМ} = 1$. Таким образом, обобщенный выигрыш при БМ и ОМ одинаков. Однако полоса частот, занимаемая сигналом при ОМ, в 2 раза меньше, чем при БМ.

Системы АМ, БМ и ОМ являются линейными, поэтому полученные выше соотношения для оптимального приемника остаются справедливыми как при слабых, так и при сильных помехах на входе. Эти соотношения определяют предельную (потенциальную) помехоустойчивость систем.

Выясним, реализуется ли эта помехоустойчивость при обычных способах приема, используемых в реальных приемниках. Для этого рассмотрим прием АМ сигналов при амплитудном (линейном) и синхронном детектировании. Перед детектором обычно включается фильтр, пропускающий частоты сигнала. Шум на выходе такого фильтра представляет собой узкополосный процесс, который можно записать согласно (2.72) в виде $N(t) = N_c(t) \cos \omega_0 t - N_s(t) \sin \omega_0 t$, где $N_c(t)$ и $N_s(t)$ — гауссовские процессы со спектральной плотностью $2N_0$ в полосе частот от 0 до $F/2$ (см. примечание на с. 53). Тогда принимаемое колебание на входе детектора будет

$$Z(t) = S(B, t) + N(t) = [u_0 (1 + m B) + N_c(t)] \cos \omega_0 t - N_s(t) \sin \omega_0 t. \qquad (6.88)$$

Огибающая этого колебания $U = \sqrt{[u_0(1+mB) + N_c(t)]^2 + N^2_s(t)}$. При слабой помехе квадратурной составляющей шума $N_s(t)$ можно пренебречь, и тогда $U \approx u_0[1 + mB(t)] + N_c(t)$.

Если в приемнике используется обычный линейный детектор огибающей, то переменная составляющая колебания на его выходе $k[mu_0 B(t) + N_c(t)]$, где k — коэффициент передачи детектора. При синхронном детектировании приемник реагирует лишь на косинусоидальную составляющую подаваемого на детектор колебания $Z(t)$ (6.88), и поэтому результат будет точен не только при слабой, но и при сильной помехе. Если усиление приемника подобрано так, что в отсутствие шума колебание его на выходе равно первичному сигналу $B(t)$, то при наличии помехи оно будет равно

$$\hat{B}(t) = B(t) + \frac{N_c(t)}{m u_0} = B(t) + \hat{N}(t).$$

Шум на выходе приемника $\dot{N}(t)=N_c(t)/(mu_0)$ является нормальным флуктуационным колебанием с нулевым средним значением. Энергетический спектр этого колебания

$$\hat{G}_{ш}(f)=\frac{2N_0}{m^2u_0^2},$$

что совпадает с выражением (6.86). Таким образом, обычный приемник с синхронным детектором реализует потенциальную помехоустойчивость при любом уровне помех на входе, а приемник с линейным детектором огибающей — только при малом уровне помех.

Можно показать, что и при детектировании сигналов БМ и ОМ синхронным детектором потенциальная помехоустойчивость реализуется при любом уровне шума на входе приемника.

Фазовая модуляция также относится к прямым системам модуляции, однако к нелинейным. При фазовой модуляции сигнал можно записать в виде: $S(B, t)=u_0\cos[\omega_0 t+\beta B(t)]$, где β — девиация фазы или индекс фазовой модуляции. Для такого сигнала $\partial S/\partial B=-\beta u_0\sin[\omega_0 t+\beta B(t)]$;

$$\overline{\left(\frac{\partial S}{\partial B}\right)^2}=\frac{1}{T}\int_0^T\overline{\left(\frac{\partial S}{\partial B}\right)^2}dt=\frac{1}{2}\beta^2u_0^2;$$

$$P_c=\frac{1}{T}\int_0^T\overline{S^2(B,t)}\,dt=\frac{1}{2}u_0^2;$$

$$G_{ш}(f)=2N_0/(\beta^2u_0^2);\quad g_{ФМ}=\beta^2\alpha/\Pi^2. \tag{6.89}$$

При больших индексах модуляции $\alpha\approx2\beta$

$$g_{ФМ}\approx\frac{\alpha^3}{4\Pi^2};\quad g'_{ФМ}\approx\frac{\alpha^2}{4\Pi^2}. \tag{6.90}$$

Как видно, при ФМ выигрыш зависит от индекса модуляции и пик-фактора сообщения. Так как β может быть больше единицы, то и выигрыш в этой системе можно получить значительно больше единицы. Платой за этот выигрыш является расширение полосы частот, занимаемой сигналом. Полученные соотношения справедливы для малого уровня шума на входе приемника, так как сигнал $S(B, t)$ при ФМ нелинейно зависит от $B(t)$.

Частотная модуляция относится к интегральным системам модуляции, потенциальная помехоустойчивость которой определяется на основании соотношения (6.78). Сигнал при ЧМ можно представить в виде $S(B, t)=u_0\cos[\omega_0 t+\Delta\omega\psi(t)]$, где $\Delta\omega$ — девиация частоты, $\psi(t)=\int_0 B(\tau)\,d\tau$. Для такого сигнала имеем

$$\frac{\partial S}{\partial\Psi}=-\Delta\omega u_0\sin[\omega_0 t+\Delta\omega\psi(t)];$$

$\overline{\left(\frac{\partial S}{\partial\psi}\right)^2}=\frac{1}{T}\int_0^T\overline{\left(\frac{\partial S}{\partial\Psi}\right)^2}dt=\frac{1}{2}\Delta\omega^2u_0^2$. Тогда на основании (6.78)

энергетический спектр шума на выходе приемника ЧМ будет

$$\hat{G}_{\text{ш}}(f)=2\left(\frac{2\pi f}{\Delta\omega}\right)^2\frac{N_0}{u_0^2}, \tag{6.91}$$

а на основании (6.80) и (6.81) получаем выражения для выигрыша:

$$g_{\text{ЧМ}}=\frac{3\Delta f F}{\Pi^3 F_c^3}=\frac{3\beta^2\alpha}{\Pi^2}\approx\frac{3}{4}\frac{\alpha^3}{\Pi^2};\quad g'_{\text{ЧМ}}=\frac{3F^2}{4\Pi^2F_c}=\frac{3\beta^2}{\Pi^2}\approx\frac{3}{4}\frac{\alpha^2}{\Pi^2}. \tag{6.92}$$

Здесь учтено, что при большом индексе модуляции $\beta=\Delta f/F_c=\Delta\omega/\Omega_c$ полоса сигнала $F\approx 2\beta F_c$. При ЧМ, также, как и при ФМ, выигрыш может быть значительно больше единицы, и достигается это за счет расширения полосы частот сигнала (увеличения индекса модуляции). Частотная и фазовая модуляция являются примерами систем, в которых верность передачи сообщений при данном уровне помех может быть повышена не только за счет увеличения мощности сигнала, как это имеет место при линейных видах модуляции, но и за счет расширения полосы частот, занимаемой сигналом. Все эти выводы и полученные выше соотношения для ЧМ, также, как и для ФМ, справедливы лишь при малом уровне помех. Эти системы, как будет показано ниже, имеют ярко выраженный пороговый эффект.

Для определения реальной помехоустойчивости при ФМ и ЧМ шум на входе детектора на основании (2.72) представим в виде

$$N(t)=U_{\text{п}}(t)\cos[\omega_0 t+\varphi(t)],$$

где

$$U_{\text{п}}(t)=\sqrt{N_c^2(t)+N_s^2(t)};\quad \varphi=\operatorname{arc\,tg}\frac{N_s(t)}{N_c(t)};$$

$$N_c(t)=U_{\text{п}}(t)\cos\varphi(t);\quad N_s(t)=U_{\text{п}}(t)\sin\varphi(t).$$

Тогда принятое колебание на входе детектора при угловой модуляции

$$Z(t)=S(B,\ t)+N(t)=u_0\cos[\omega_0 t+\psi(t)]+$$
$$+U_{\text{п}}(t)\cos[\omega_0 t+\varphi(t)]=U(t)\cos[\omega_0 t+\theta(t)].$$

Векториая диаграмма этого колебания представлена на рис. 6.16.

Если в качестве опорной величины выбрать фазовый угол сигнала $\psi(t)$, то мгновениая фаза сигнала $Z(t)$ может быть представлена выражением

$$\theta(t)=\psi(t)+\operatorname{arc\,tg}\frac{U_{\text{п}}(t)\sin(\varphi-\psi)}{u_0+U_{\text{п}}\cos(\varphi-\psi)}.$$

При слабом шуме, когда $U_{\text{п}}(t)\ll u_0$,

$$\theta(t)\approx\psi(t)+\frac{U_{\text{п}}(t)}{u_0}\sin(\varphi-\psi).$$

Рассмотрим для простоты момент времени, когда $|\psi|(t)|\ll 1$. В этом случае при малой помехе

$$\theta(t) \approx \psi(t) + U_{\text{п}} \sin\varphi/u_0 = \psi(t) + N_s(t)/u_0. \qquad (6.93)$$

К тем же результатам легко прийти и на основании векторной диаграммы (рис. 6.10). Из рисунка непосредственно следует, что $\theta(t) = \psi(t) + \gamma(t)$, где $\operatorname{tg}\gamma = N_s(t)/[u_0 + N_c(t)]$. При слабом шуме ($|\dot{N}| \ll u_0$) $\operatorname{tg}\gamma \approx \gamma = N_s(t)/u_0$ и $\theta(t) = \psi(t) + \dfrac{N_s(t)}{u_0}$. В случае фазовой модуляции $\psi(t) = \beta B(t)$ и выражение (6.93) принимает вид $\theta(t) \approx \beta B(t) + N_s(t)/u_0$. При соответствующем подборе коэффициента усиления приемника колебание на выходе фазового детектора можно представить в виде $\hat{B}(t) = B(t) + N_s(t)/(\beta u_0) = B(t) + \hat{N}(t)$, где $\hat{N}(t) = N_s(t)/(\beta u_0)$ представляет собой нормальный шум на выходе приемника. Энергетический спектр этого шума $G_{\text{ш}}(f) = 2N_0/(\beta^2 u^2_0)$ совпадает с выражением (6.89) для ФМ.

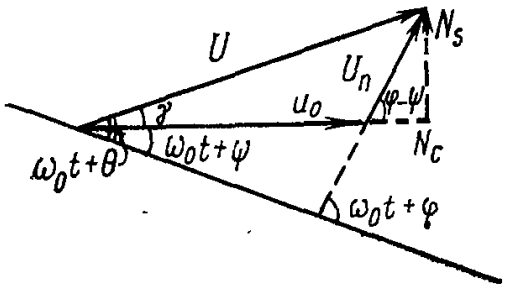


Рис. 6.10. Векторная диаграмма суммы сигнала и узкополосного шума

В случае частотной модуляции $\psi(t) = \Delta\omega \int_0^t B(t)dt$, и выражение (6.93) принимает вид

$$\theta(t) \approx \Delta\omega \int_0^t B(\tau)\, d\tau + N_s(t)/u_0.$$

Частотный детектор реагирует на производную фазы θ' принимаемого сигнала $\theta' = \Delta\omega B(t) + N'_s(t)/u_0$. Отсюда колебание на выходе частотного детектора можно представить как

$$\hat{B}(t) = B(t) + N'_s(t)/(\Delta\omega u_0) = B(t) + \hat{N}(t).$$

Шум на выходе приемника $\hat{N}(t) = N'_s(t)/(\Delta\omega u_0)$ является нормальным флуктуационным колебанием со средним значением, равным нулю. Энергетический спектр этого шума определяется как спектр производной: $G_{\text{ш}}(f) = (2\pi f)^2 2N_0/(\Delta\omega^2 u^2_0)$, что совпадает с выражением (6.91). Таким образом, при малом уровне шума обычные способы приема ФМ и ЧМ сигналов реализуют потенциальную помехоустойчивость.

6.7. ПОРОГ ПОМЕХОУСТОЙЧИВОСТИ

Все широкополосные системы модуляции обеспечивают высокую помехоустойчивость при условии, что отношение сигнала к помехе $\rho_{\text{вх}}$ на входе приемника больше некоторого предельного (порогового) значения $\rho_{\text{пр}}$. При $\rho_{\text{вх}} < \rho_{\text{пр}}$ широкополосные системы теряют свои преимущества (резко снижается помехоустойчивость) и связь становится практически невозможной. Значение порога определяет предельную дальность связи при заданной мощности передатчика. Поэтому важно определить это значение и установить закон изменения отношения сигнала к помехе на выходе приемника за порогом (при $\rho_{\text{вх}} < \rho_{\text{пр}}$).

Рассмотрим, как объясняется пороговый эффект теорией информации.

Обратимся вновь к выражению (6.81) для обобщенного выигрыша, выразив его в децибелах:

$$g'_{[дБ]} = (\rho_{вых})_{[дВ]} - (\rho_{вх}\,\alpha)_{[дБ]}. \qquad (6.94)$$

Заметим, что $\rho_{вх}\alpha = \dfrac{P_c}{(N_0 F)}\, F/F_c = \dfrac{P_c}{N_0 F_c} = \rho'_{вх}$ есть отношение мощности сигнала на входе приемника к мощности той части шума, которая лежит в полосе частот, равной ширине спектра сообщения F_c (а не в полосе спектра сигнала F). Равенство F_c и F имеет место при непосредственной передаче, а также при однополосной модуляции.

Если для некоторой системы связи построена зависимость $\rho_{вых}$ от $\rho'_{вх}$ в децибелах, то для любого значения $\rho_{вых}$ обобщенный выигрыш $g'_{[дВ]}$ получается путем измерения разности ординат данной кривой и прямой, проведенной под углом 45° (рис. 6.11). В идеальной системе согласно (6.84)

$$\lg \rho_{вых} = \alpha \lg (1 + \rho_{вх}) \qquad (6.95)$$

или для случая, когда $\rho_{вх} \gg 1$, $\lg \rho_{вых} \approx \alpha \lg \rho_{вх} = \alpha \lg \dfrac{\rho'_{вх}}{\alpha}$,

$$\rho_{вых\,[дБ]} = \alpha\,(\rho'_{вх\,[дБ]} - \alpha_{[дБ]}). \qquad (6.96)$$

Это не что иное, как семейство прямых. На рис. 6.11 они изображены сплошными линиями. Для $\alpha = 1$ прямая проходит через начало координат под углом 45°; для $\alpha > 1$ эти прямые идут круче и пересекают ось абсцисс правее начала координат. Из (6.96) получаем обобщенный выигрыш (в децибелах) для идеальных систем, вычитая из него $\rho'_{вх\,[дБ]}$:

$$g'_{макс\,[дБ]} = (\alpha - 1)\,(\rho'_{вх\,[дБ]} - \alpha_{[дБ]}). \qquad (6.97)$$

Из выражения (6.97) следует, что в идеальной системе обобщенный выигрыш в логарифмическом масштабе растет линейно с увеличением $\rho'_{вх}$. Как было показано в § 6.6, для реальных систем обобщенный выигрыш g' не зависит от $\rho_{вх}$, если $\rho_{вх} \gg 1$. Тогда соответствующие им зависимости на графике рис. 6.3 представляют собой прямые, параллельные прямой, проходящей через начало координат под углом 45°, но сдвинутые на величину g' (на рис. 6.11 одна из таких кривых изображена пунктирной линией). Если бы кривые для реальных и идеальных систем пересекались, то это бы означало, что реальная система ста-

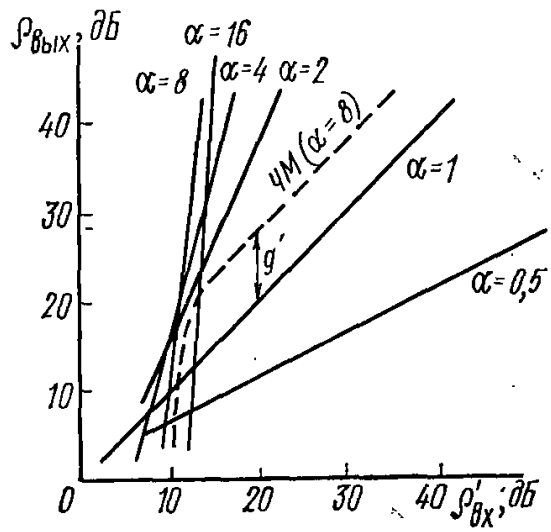


Рис. 6.11. Зависимость $\rho_{вых}$ от $\rho'_{вх}$ для идеальной системы связи

ла лучше идеальной. Очевидно, что это невозможно и кривые для реальных систем располагаются всегда не выше соответствующих кривых для идеальных систем. Отсюда можно сделать вывод, что системы, дающие выигрыш при больших $\rho_{вх}$, не могут сохранять его постоянным с уменьшением $\rho_{вх}$. Начиная с некоторого порогового значения этого отношения $\rho_{вх}$, выигрыш системы резко уменьшается. Более того, выигрыш может стать отрицательным (проигрышем), если $\rho'_{вх}<\alpha$. Чем больше α, тем сильнее сказывается пороговый эффект. Системы с $\alpha\leqslant 1$ вообще не подвержены пороговым явлениям, но они и при больших $\rho_{вх}$ не дают выигрыша. Зато они позволяют передавать сообщения по каналу с полосой пропускания, равной или меньшей ширины спектра сообщения. Для этого сообщение должно быть преобразовано в сигнал, спектр которого уже спектра сообщения. Чтобы получить высокую верность передачи в такой системе, необходимо иметь достаточно большое отношение мощности сигнала к мощности помехи в канале, поскольку при $\alpha<1$ $\rho_{вых}<\rho_{вх}$. На практике системы с $\alpha<1$ применения пока не находят.

Поясним теперь явление порога геометрически. Для этого воспользуемся понятиями пространства сообщений и пространства сигналов, введенными в § 2.5. Каждому сигналу $s(t)$ соответствует точка в n-мерном пространстве, где $n=2TF$ — база сигнала, а ансамблю возможных сигналов — некоторая область в этом пространстве сигналов. При фиксированной мощности область возможных сигналов представляет собой гиперсферу с радиусом $r_c=\sqrt{nP_c}=\sqrt{2TFP_c}$, а область принятых колебаний $Z(t)=S(t)+N(t)$ — сферу с радиусом $r_z=\sqrt{2TF(P_c+P_ш)}$. Вокруг каждой точки переданного сигнала образуется область неопределенности, обусловленная помехой. Если помехой является гауссовский шум, то эта область имеет сферическую форму с радиусом $r_ш=\sqrt{2TFP_ш}$.

Как уже отмечалось, модуляция является отображением пространства сообщений на пространство сигналов, а демодуляция — обратным отображением пространства принятых сигналов в пространство принятых сообщений (оценок). В общем случае размерность (база) пространства сообщений $2F_cT$ отличается от размерности пространства сигналов $2FT$.

Рассмотрим некоторый отрезок прямой в области пространства возможных сообщений (эта область определяется, например, информацией сообщения). При модуляции этот отрезок отобразится в некоторую линию в пространстве сигналов, каждая точка s которой соответствует определенной реализации сообщения. Форма и длина этой линии зависят от вида модуляции. Помеха, наложившаяся на сигнал, вызовет смещение точки s в некоторую точку $\hat{s}=s+\Delta s$, соответствующую другому сообщению $\hat{b}=b+\Delta b$, аналогично тому, как было показано на рис. 6.6. Очевидно, чем больше $\Delta s/\Delta b$, тем меньше ошибка в принятом сообщении при той же помехе.

Для увеличения помехоустойчивости (отношения $\Delta s/\Delta b$) необходимо увеличивать длину линии сигналов, соответствующей данному отрезку в пространстве сообщений. Это можно сделать, расширяя используемую область сигналов (радиус r_c). Но поскольку мощность сигнала ограничена, для этого нужно увеличивать базу сигнала, расширяя его спектр. При этом линия сигнала может приобрести сложную извилистую форму.

На рис. 6.12 показан пример такой линии сигналов, причем для наглядности использовано двумерное пространство сигналов (при одномерном пространстве сообщений). Заметим, что при линейной модуляции линии сигнала и сообщений подобны и поэтому выигрыш в них невозможен. В случае широкополосных систем (например, ЧМ) линия сигналов имеет сложную извилистую форму. Длина линии увеличивается с увеличением полосы частот сигна-

ла. Однако при ограниченной мощности сигнала удлинение линии ведет к сближению различных витков линии. Слабая помеха сдвигает точку принимаемого колебания вдоль линии сигнала. Эти сдвиги определяют величину нормальной ошибки. Сильная помеха может перевести точку принимаемого колебания на другой «виток» линии сигнала. Эти «перескоки» и определяют аномальные ошибки.

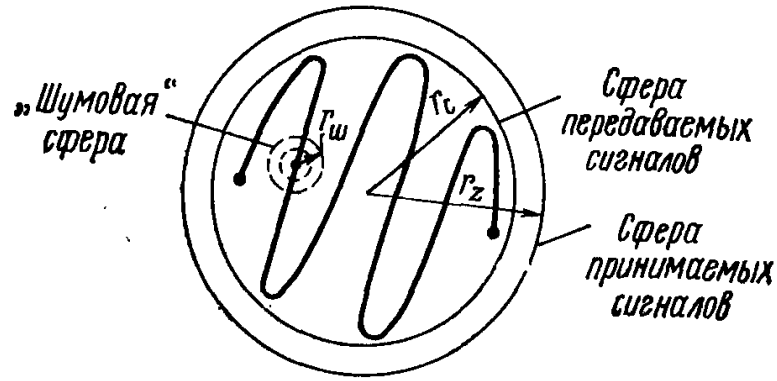


Рис. 6.12. К геометрической интерпретации порога

Более точная теория [6], в которой при вычислении пропускной способности канала и скорости передачи информации учитывается, что сигнал не является гауссовским, позволяет найти пороговое значение $\rho_{вх}=\rho_{пр}$, а также вычислить значения $\rho_{вых}$ ниже порога.

На рис. 6.13 приведена вычисленная зависимость отношения сигнала к помехе на выходе приемника ЧМ $\rho_{вых} = f(\rho'_{вх})$ при различных индексах мо-

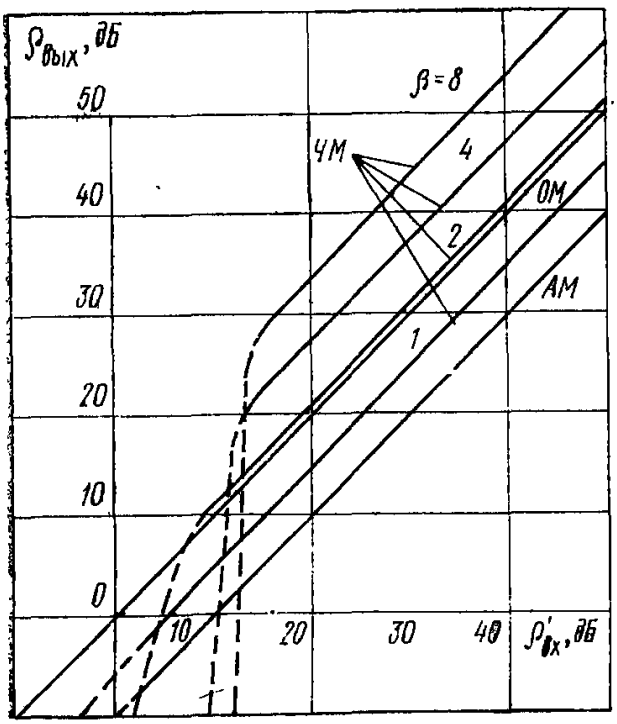


Рис. 6.13. Кривые верности для ЧМ ($\beta=1, 2, 4, 8$), АМ и ОМ

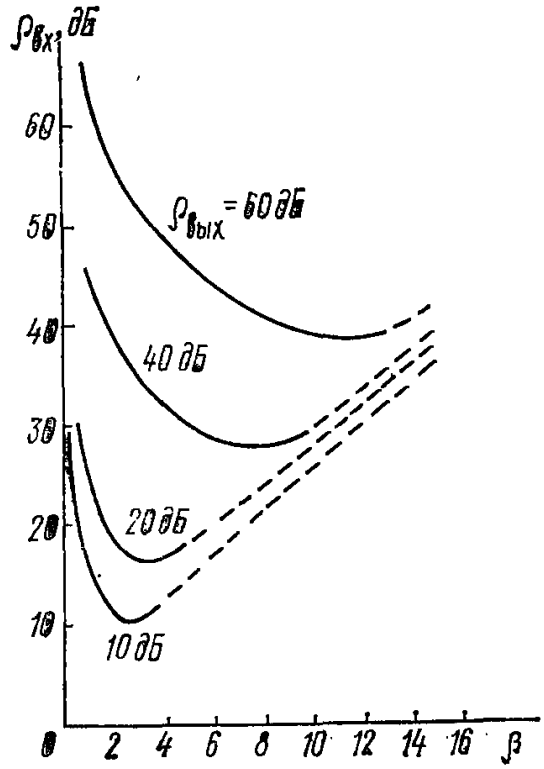


Рис. 6.14. Зависимость необходимой мощности сигнала для заданного качества приема от индекса модуляции при ЧМ

дуляцнн н пнкфакторе П=3. Здесь по осн абсцнсс отложено отношенне $\rho'_{вх}$ снгнала к помехе в децнбелах на входе прнемнпка в полосе F_c, а по осн орднпат — отношенне снгнала к помехе в децибелах на выходе прнемнпка ЧМ. Здесь же, для сравненпя, построены крнвые для АМ н ОМ.

Прнведенные на рнс. 6.13 крнвые показывают, что с увелнченнем индекса модуляцнн пороговое значенне $\rho'_{пр}$ увелнчнвается н порог становится более резко выраженным. За порогом помехоустойчнвость ЧМ снльно уменьшается н прн сравнптельно большом уровне помех становптся хуже, чем помехоустойчнвость снстем ОМ н АМ. Прн $\beta \leqslant 1$ порог выражен слабо н крнвые для ЧМ прнблнжаются к крнвой АМ. Порог прн ОМ отсутствует.

Полученное в теорнн пороговое значенне отношенпя снгнал/помеха на входе прнемпика $\rho_{пр}=\rho_{пр}/\alpha$ с увелнченнем нндекса модуляцнн нзменяется в небольшнх пределах. Пороговая мощность $P_{с.пр}$ прн этом существенно возрастает, поскольку

$$P_{с.пр} = \rho_{пр} P_{ш}, \tag{6.98}$$

где $P_{ш}=N_0F$ — мощность шума на входе прнемннка, пропорцнональная шнрнне полосы пропускання прнемнпка F н, следовательно, нндексу модуляцнн β.

При фиксированной спектральной плотности шума N_0 величина $\rho'_{вх}$ пропорциональна мощности сигнала. На рис. 6.14 показана расчетная зависимость $\rho'_{вх}$ от индекса модуляции β при заданной верности приема $\rho_{вых}$. Пунктирные части кривых соответствуют условиям ниже порога помехоустойчивости. Из этого рисунка видно, что существует оптимальное значение индекса модуляции, при котором для обеспечения заданной верности приема затрачивается минимальная мощность сигнала. Это значение индекса тем больше, чем выше требуемая верность приема, примерно совпадает с пороговой точкой.

Полученные теоретические значения порога для различных видов модуляции являются предельными, которые в реальных приемниках могут быть достигнуты, но не могут быть превзойдены. По экспериментальным данным порог в системе ЧМ при обычном способе приема наступает примерно при равенстве пиковых значений сигнала и помехи, что соответствует $\rho_{вх}$ существенно выше теоретического. Это означает, что при большом уровне помех реальная помехоустойчивость приемника ЧМ значительно меньше потенциальной. Следовательно, имеется возможность путем усовершенствования схемы приемника снизить порог помехоустойчивости и тем самым увеличить дальность связи при той же мощности передатчика. Эта проблема стала особенно актуальна в связи с разработкой сверхдальних космических линий связи.

Большинство методов снижения порога при ЧМ основано на принципе «сжатия» спектра сигнала в приемнике. Ширина полосы пропускания обычного ЧМ приемника выбирается такой, чтобы вмещать весь эффективный спектр полезного сигнала. Эта полоса намного превышает ширину спектра передаваемого сообщения. Однако если рассмотреть ЧМ сигнал на коротком отрезке времени, то он не занимает всю эту полосу. Можно лишь утверждать, что в любой момент времени он находится в некотором отрезке полосы. Наличие априорной информации о скорости изменения частоты сигнала позволяет осуществить слежение за этой

частотой узкополосным приемником. Это достигается введением обратной связи по частоте или применением следящего фильтра промежуточной частоты. Предельное значение порога $\rho_{пр}$ может быть достигнуто лишь в схеме оптимального приемника (см. § 6.10).

6.8. ПОМЕХОУСТОЙЧИВОСТЬ СИСТЕМ С ПОДНЕСУЩИМИ

В технике связи нередко применяется двойная модуляция. При этом передаваемым сообщением (первичным сигналом) модулируется вспомогательное колебание (поднесущая) с частотой $f_{пп}$, а затем полученным модулированным колебанием модулируется другое колебание (несущая) с частотой $f_н \gg f_{пн}$. В многоканальных системах двойная модуляция позволяет осуществить частотное уплотнение каналов.

В одноканальных системах с поднесущей двойная модуляция позволяет снизить требования к стабильности несущей частоты, заменив его более легко выполнимым требованием к стабильности поднесущей частоты. Кроме того, такие системы, как ЧМ-АМ или ФМ-АМ, применяются в тех случаях, когда нельзя использовать непосредственно ЧМ или ФМ из-за селективных замираний. Так, например, в телеметрических системах необходимо передавать не только форму, но и масштаб сообщения $B(t)$. Для этих целей нельзя в каналах с замиранием использовать АМ или ОМ, так как замирания будут вызывать изменения амплитуды принимаемого сообщения. В таких случаях часто используют ЧМ-АМ или ФМ-АМ. По этой же причине система ЧМ-АМ применяется в фототелеграфных системах на коротких волнах.

Разнообразие систем модуляции позволяет осуществить большое количество систем с поднесущими. Практическое применение находят системы ОМ-АМ, ФМ-АМ, ЧМ-АМ, ЧМ-ОМ, ОМ-ЧМ, ЧМ-ЧМ и др.

При малом уровне помех выигрыш систем с поднесущими определяется таким же методом, что и при обычной однократной модуляции с помощью формул (6.80) и (6.81). Рассмотрим для примера систему ФМ-АМ. В этой системе сигнал можно выразить так:

$$S(t) = u_0 \{1 + m_н \cos[\omega_{пн} t + \beta_{пи} B(t)]\} \cos(\omega_н t + \varphi). \qquad (6.99)$$

Так как система ФМ-АМ относится к прямым системам модуляции, то выигрыш можно определить по формулам (6.80) и (6.81) с учетом выражения (6.77).

Для сигнала (6.99) имеем

$$\overline{\left(\frac{\partial S}{\partial B}\right)^2} = \frac{1}{2} u_0^2 m_и^2 \beta_{пи}^2;$$

$$P_с = \overline{S^2(t)} = \frac{1}{2} u_0^2 \left(1 + \frac{m_и^2}{П^2}\right).$$

Тогда обобщенный выигрыш в системе ФМ-АМ

$$g' = \frac{\beta_{пи}^2}{П^2} \frac{m_н^2}{2 + m_н^2} = g'_{ФМ} g'_{АМ}. \qquad (6.100)$$

Из полученного выражения следует, что обобщенный выигрыш равен произведению выигрышей при ФМ и АМ. Легко убедиться, что для любой системы с поднесущей, в которой модуляция несущей является прямой, обобщенный выигрыш g' равен произведению выигрышей первой $g'_{пн}$ и второй $g'_н$ ступеней модуляции, т. е.

$$g' = g'_{пи} g'_и. \qquad (6.101)$$

Порог помехоустойчивости в системах с поднесущими в общем случае будет наблюдаться по обеим ступеням модуляции. Здесь важно определить, по какой ступени модуляции порог наступает раньше и где он выражен сильнее. Если в одной ступени модуляции используется ОМ, БМ или АМ, а в другой ступени ФМ или ЧМ, то порог помехоустойчивости определяется той ступенью, в которой применена фазовая или частотная модуляция. Так, в системах ЧМ-АМ порог определяется модуляцией поднесущей, а в системе ОМ-ЧМ — модуляцией несущей.

6.9. ПОМЕХОУСТОЙЧИВОСТЬ СИСТЕМ С ИМПУЛЬСНОЙ МОДУЛЯЦИЕЙ

Согласно теореме Котельникова непрерывное колебание $B(t)$ может быть передано по линии связи с необходимой точностью путем передачи отдельных мгновенных значений этого колебания:

$$\ldots B(-2T_i),\quad B(-T_i),\quad B(T_i),\quad B(2T_i)\ldots, \tag{6.102}$$

взятых для моментов времени, отстоящих друг от друга на величину $T_i \leqslant 1/2F_c$, где F_c наивысшая частота, содержащаяся в колебании $B(t)$.

В системах связи, основанных на этом принципе (импульсных системах), для передачи колебания $B(t)$ используется периодическая последовательность импульсов

$$f(t) = \sum_{\kappa=-\infty}^{\infty} v(t-\kappa T_i). \tag{6.103}$$

При этом один из параметров этой последовательности изменяется в соответствии с изменением мгновенных значений (6.102) передаваемого колебания $B(t)$.

Модулированную последовательность импульсов на выходе первой ступени модуляции можно записать в виде

$$f(B,\ t) = \sum_{\kappa=-\infty}^{\infty} v[B(\kappa T_i),\quad t-\kappa T_i]. \tag{6.104}$$

Здесь $v(t)$ — огибающая импульса, причем $v(t)=0$ при $t<0$ и $t>\tau_0$, где τ_0 — длительность импульса.

Для передачи колебания $f(B,\ t)$ по радио необходимо применить еще одну ступень модуляции. При этом может быть использована любая из систем модуляции, рассмотренных в § 6.6. Однако чаще всего во второй ступени модуляции модулированная последовательность импульсов $f(B,\ t)$ перемножается с гармоническим напряжением несущей, в результате чего получается сигнал

$$S(B,\ t) = f(B,\ t)\, a\cos\omega_н t =$$

$$= a\sum_{\kappa=-\infty}^{\infty} v[B(\kappa T_i),\quad t-\kappa T_i]\cos\omega_н t. \tag{6.105}$$

Эту операцию перемножения называют амплитудной модуляцией и обозначают соответствующие системы сокращениями типа АИМ-АМ, ФИМ-АМ, ШИМ-АМ и т. д.

Особенностью такого вида вторичной модуляции является то, что при нем сохраняется одно из преимуществ импульсных систем связи — наличие пауз между короткими отрезками сигнала. Эти паузы, в частности, можно использовать для уплотнения канала (см. гл. 8).

Сигнал, излучаемый передатчиком, представляет собой колебание, зависящее от времени и передаваемого сообщения:

$$S(B,\ t)=S[B(\kappa' T_t),\ t]. \tag{6.106}$$

Потенциальную помехоустойчивость импульсных систем модуляции можно определить исходя из общей формулы (6.77) для прямых систем модуляции несущей или формулы (6.78) для интегральных систем.

Подробные исследования помехоустойчивости импульсных систем модуляции читатель может найти в работе [6]. Здесь мы ограничимся лишь изложением результатов.

Потенциальная помехоустойчивость системы АИМ-АМ оказывается равной потенциальной помехоустойчивости аналоговой системы АМ и, следовательно, очень низкой.

Из всех систем импульсной модуляции наиболее помехоустойчивыми являются ФИМ и ЧИМ. Так как первая из них несколько проще, то для передачи сигналов по каналам связи используется почти исключительно ФИМ-АМ. Виды модуляции АИМ и ШИМ в каналах связи практически не используются и применяются лишь в процессе обработки сигналов.

Остановимся поэтому подробнее только на определении помехоустойчивости системы ФИМ-АМ.

Обобщенный выигрыш в системе ФИМ-АМ определяется следующим выражением:

$$g^{\prime}=\frac{1}{\pi^2\kappa_{\Phi}^2}\left(\frac{\Delta\tau}{\tau_0}\right)^2, \tag{6.107}$$

где κ_{Φ} — коэффициент зависящий от формы импульса; $\Delta\tau$ — максимальная девиация положения импульса.

Как видим, помехоустойчивость системы ФИМ зависит от формы импульса. При оптимальном выборе параметров системы ФИМ-АМ: $F\tau_0\approx 1$, $f_i=2F_c$, $\Delta\tau\approx 1\,(/2f_i)$ —

$$g'_{op}t=\frac{\alpha^2}{16\,\pi^2\kappa_{\Phi}^2}. \tag{6.108}$$

При $\kappa^2{}_{\Phi}=\frac{1}{12}$, что соответствует треугольной форме импульсов,

$$g'_{opt}=\frac{3}{4}\frac{\alpha^2}{\pi^2}. \tag{6.109}$$

Это совпадает с формулой (6.92) для системы с частотной модуляцией. Следовательно, потенциальная помехоустойчивость системы ФИМ-АМ при оптимальном выборе параметров и системы ЧМ при том же значении α одинакова. Впрочем, как мы скоро

убедимся, при ФИМ можно использовать более широкий спектр, чем при ЧМ, т. е. иметь большее значение α и больший обобщенный выигрыш.

Формулы, определяющие интенсивность шума на выходе приемника в системе с ФИМ, справедливы и для соответствующих систем с ШИМ. Однако при этом следует помнить, что в системе с ШИМ средняя длительность импульсов больше, чем в системе с ФИМ. Поэтому потенциальная помехоустойчивость ШИМ ниже, чем ФИМ.

В настоящее время широко применяется пороговый прием сигналов с ФИМ-АМ, при котором момент прихода импульса определяется как момент перехода огибающей через некоторый пороговый уровень. Относительно слабые помехи слегка изменяют форму огибающей и поэтому смещают момент пересечения порога. Чем круче фронт импульса, тем меньше смещение, вызываемое помехой. Поэтому расширение спектра, позволяющее увеличить крутизну фронта, приводит при слабых помехах к увеличению выигрыша (6.108). При большом уровне помех отдельные их выбросы могут превысить пороговый уровень и вызывать ложные срабатывания выходных устройств в промежутках между импульсами. Кроме этого, выбросы помехи могут подавлять отдельные импульсы сигнала. Если пиковое значение помехи равно или больше порогового уровня приемника, то вероятность ложных срабатываний (аномальных ошибок) резко увеличивается и прием становится практически невозможным.

Таким образом, порог помехоустойчивости определяется порогом срабатывания приемника. Обычно пороговый уровень в приемнике выбирается равным половине амплитуды сигнала в пике импульса. В этом случае порог помехоустойчивости при пикфакторе помехи $\Pi = 4$ будет

$$\rho_{пр} \approx 32 \frac{\tau_0}{T_i}. \tag{6.110}$$

Пороговая мощность при этом равна

$$P_{пр} \approx 32 \frac{\tau_0}{T_i} N_0 F. \tag{6.111}$$

При оптимальном выборе параметров системы $F\tau_0 = 1$ и $T_i = \dfrac{1}{2F_c}$

$$P_{пр} \approx 64 N_0 F_c. \tag{6.112}$$

Полученные выражения показывают, что пороговая мощность сигнала в импульсных системах с амплитудной модуляцией несущей в первом приближении не зависит от полосы пропускания приемника[1]. Это обусловливает одно из важнейших преимуществ импульсных систем по сравнению, например, с частотной модуляцией. Действительно, выбирая достаточно малую длительность импульса, можно увеличить величину выигрыша системы, не увеличивая при этом пороговую мощность сигнала. В системе же с частотной модуляцией увеличение выигрыша за счет расширения полосы частот канала ведет к увеличению пороговой мощности сигнала. Именно поэтому при ФИМ-АМ можно реализовать большее значение α, чем при ЧМ.

6.10. ТЕОРИЯ НЕЛИНЕЙНОЙ ФИЛЬТРАЦИИ

В теории оптимального приема В. А. Котельникова в качестве модели сообщения рассматривались либо случайная равномерно распределенная величи-

[1] Постоянство пороговой мощности объясняется тем, что при расширении спектра можно укоротить импульс и увеличить его амплитуду при неизменной мощности сигнала, а, следовательно, и повысить порог срабатывания.

на (прн оценке непрерывного параметра), лнбо стацнонарный случайный процесс с огранпченным спектром (прн оценке непрерывных колебанпй). Такпе модели представляют собой снльную идеализацию реальных сообщенпй [16].

В работах Р. Л. Стратоновпча, В. И. Тихонова и других развита теория, в которой широко используются другие модели сообщенпй. В настоящее время эта теория, получившая название теории нелинейной фильтрации, наиболее разработана для двух случаев: 1) когда передаваемое сообщение $B(t)$ является нормальным случайным процессом и 2) когда сообщение $B(t)$ представляет собой марковский процесс. Мы остановимся лишь на теории фильтрации одномерных марковских гауссовских процессов.

Реальные сообщения можно приближенно моделировать марковским процессом[1], который описывается следующим стохастическим дифференциальным уравнением:

$$\frac{dB}{dt} + \alpha B = V(t), \tag{6.113}$$

где $V(t)$ — стационарный белый шум с известными статистическими характеристиками:

$$\overline{V(t)} = 0, \quad \overline{V(t_1)\,V(t_2)} = \frac{1}{2} G_V \delta(t_2 - t_1). \tag{6.114}$$

Здесь G_V и α — известные постоянные коэффициенты. Такое сообщение мы уже рассматривали в § 6.2 при обсуждении фильтра Калмана. Различие заключается лишь в том, что теперь рассматривается не непосредственная передача сообщения $B(t)=S(t)$ по гауссовскому каналу, а передача модулированного сигнала $S(B(t), t)$.

Не нарушая общности, можно положить $G_V = 4\alpha$. Тогда $B(t)$ будет нормированным безразмерным процессом с единичной дисперсией (мощностью), а изменения глубины модуляции того или иного параметра скажутся лишь на коэффициенте (индексе) модуляции.

Поскольку процесс, описываемый стохастическим уравнением (6.113), является марковским (диффузионным), изменения во времени его плотности вероятности $w(b, t)$ определяются уравнением Колмогорова — Фокера — Планка (2.30), которое в данном случае имеет вид

$$\frac{\partial}{\partial t} w(b, t) = -\frac{\partial}{\partial b}[\alpha w(b, t)] + \frac{1}{4} G_V \frac{\partial^2}{\partial b^2} w(b, t). \tag{6.115}$$

Принимаемое колебание $Z(t)$ на некотором интервале $(0, T)$ представляет собой сумму сигнала $S(B(t), t)$, зависящего от передаваемого сообщения $B(t)$, и стационарного белого шума $N(t)$:

$$Z(t) = S(B(t), t) + N(t), \quad 0 \leqslant t \leqslant T. \tag{6.116}$$

Предполагается, что характеристики шума известны:

$$\overline{N(t)} = 0; \quad \overline{N(t_1)\,N(t_2)} = \frac{1}{2} N_0 \delta(t_2 - t_1).$$

Для большей общности будем рассматривать передачу сигнала в канале с флуктуирующей фазой $\varphi(t)$, которую будем представлять нестационарным марковским процессом с независимыми приращениями, описываемым дифференциальным уравнением

$$\frac{d\varphi}{dt} = W(t), \tag{6.117}$$

---

[1] Реальные сообщения, как правило, не являются марковскими, поскольку зависимость между их отдельными сечениями более сложная. Тем не менее марковская аппроксимация описывает реальное сообщение лучше, чем аппроксимация процессом с финитным спектром, у которого сечения, разделенные котельниковским интервалом, вовсе не коррелированны.

где $W(t)$ — центрированный белый шум с корреляционной функцией $B_\varphi(t_1, t_2)=\frac{1}{2} G_W \delta(t_2-t_1)$. Все белые шумы, фигурирующие в данной задаче, взаимно независимы.

Располагая этими априорными данными, нужно найти устройство, которое бы с наилучшей точностью воспроизводило изменяющееся во времени случайное колебание $B(t)$. Вся доступная информация о передаваемом сообщении $B(t)$ содержится в апостериорной плотности вероятности $w(b|z)$. Поэтому первый этап решения задачи состоит в определении этой апостериорной плотности вероятности с учетом результата наблюдения принятой реализации $z(t)$ в предшествующий промежуток времени. В работах Р. Л. Стратоновича показано, что апостериорная плотность вероятности $w(b|z)$ реализации $b(t)$ в конечный момент времени наблюдения определяется следующим нелинейным дифференциальным уравнением относительно апостериорной плотности вероятности $w(b|z)$ [15]:

$$\frac{\partial}{\partial t} w(b|z) = -\frac{\partial}{\partial b}[A_1(b, t) w(b|z)] + \\ + \frac{1}{2}\frac{\partial^2}{\partial b^2}[A_2(b, t) w(b|z)] + [F(b, t) - \overline{F(b, t)}] w(b|z). \quad (6.118)$$

Здесь A_1 и A_2 — коэффициенты сноса и диффузии; $F(b, t)$ — производная по времени от логарифма функции правдоподобия:

$$F(b, t) = \frac{d}{dt} \ln w(z|b). \quad (6.119)$$

Уравнение (6.118) определяет полную процедуру фильтрации сообщения на фоне белого шума. Оптимальное приемное устройство должно моделировать уравнение (6.118) и определять оценку $\hat{b}(t)$, соответствующую максимуму апостериорной вероятности. В общем случае аналитическое решение этого уравнения оказывается трудной задачей. Схемы оптимального приемника при этом получаются весьма сложными. Для получения более простых схем оптимальных устройств целесообразно использовать приближенные решения. При достаточно больших отношениях сигнал/шум и большом времени наблюдения есть основания считать, что апостериорная плотность вероятности будет приближенно нормальной:

$$w(b|z) \approx \frac{1}{\sqrt{2\pi\sigma_b^2(t)}} e^{-\frac{[b(t)-\hat{b}(t)]^2}{2\sigma_b^2(t)}}. \quad (6.120)$$

Такую аппроксимацию $w(b|z)$ в теории нелинейной фильтрации принято называть гауссовским приближением, а получающийся при этом алгоритм обработки колебания $Z(t)$ — квазиоптимальным. Решение задачи в этом случае существенно упрощается, так как апостериорная плотность вероятности (6.110) определяется всего двумя параметрами: средним значением $\hat{b}(t)$, определяющим оптимальную оценку передаваемого сообщения $b(t)$, и дисперсией $\sigma^2_b(t)$, характеризующей ошибку фильтрации.

Если считать коэффициент сноса $A_1(b, t)$ линейной функцией b, а коэффициент диффузии $A_2(b, t)=\text{const}$, то после подстановки (6.120) в (6.118) и несложных вычислений получим для определения $\hat{b}(t)$ и $\sigma^2_b(t)=\kappa(t)$ систему из двух обыкновенных нелинейных дифференциальных уравнений:

$$\left.\begin{aligned} \frac{d}{dt}\hat{b}(t) &= A_1(\hat{b}, t) + \kappa(t)\frac{\partial F(\hat{b}, t)}{\partial \hat{b}}; \\ \frac{d}{dt}\kappa(t) &= 2\kappa(t)\frac{\partial A_1(\hat{b}, t)}{\partial \hat{b}} + \kappa^2(t)\frac{\partial^2 F(\hat{b}, t)}{\partial \hat{b}^2} + A_2. \end{aligned}\right\} \quad (6.121)$$

Таким образом, задача оптимальной фильтрации в гауссовом приближении сводится к совместному решению или моделированию уравнений (6.121). Система, моделирующая эти уравнения (нелинейный фильтр), будет воспроизводить переданное сообщение $b(t)$ с минимальной среднеквадратической ошибкой (по крайней мере, в случае слабых помех, когда оправдано гауссово приближение).

Для марковского гауссовского процесса, описываемого стохастическим дифференциальным уравнением (6.113),

$$A_1(\hat{b}, t) = -\alpha b(t), \quad A_2(\hat{b}, t) = \frac{1}{2} G_V,$$

уравнения (6.121) запишутся так:

$$\left.\begin{aligned} \frac{d\hat{b}(t)}{dt} &= -\alpha \hat{b}(t) + \kappa(t) \frac{\partial F(\hat{b}, t)}{\partial \hat{b}}; \\ \frac{d}{dt}\kappa(t) &= -2\alpha\kappa(t) + \frac{1}{2} G_V + \kappa^2(t) \frac{\partial^2 F(\hat{b}, t)}{\partial \hat{b}}. \end{aligned}\right\} \qquad (6.122)$$

Если сигнал $S(t, B)$ линейно зависит от передаваемого сообщения $B(t)$, то уравнения (6.122) вырождаются в уравнения (6.27), (6.32), описывающие линейный оптимальный фильтр Калмана.

Рассмотренная теория нелинейной фильтрации может быть обобщена и на случай, когда передаваемое сообщение описывается негауссовским марковским процессом. В этом случае вместо уравнения (6.113) фигурирует нелинейное стохастическое дифференциальное уравнение.

Другим обобщением является передача сигнала, модулированного сообщением $B_1(t)$, которое можно аппроксимировать компонентой многомерного марковского процесса (см. сноску на с. 204).

В качестве примера рассмотрим применение теории нелинейной фильтрации для синтеза оптимальных приемников в системах с фазовой и частотной модуляцией.

Сигнал при фазовой модуляции имеет вид

$$S(B, t) = u_0 \cos[\omega_0 t + \theta(t)], \quad \theta(t) = \beta B(t) + \varphi(t), \qquad (6.123)$$

где u_0 и ω_0 — априорно известные значения амплитуды и частоты; β — индекс фазовой модуляции; $\varphi(t)$ — процесс, описывающий флуктуацию начальной фазы; $B_1(t)$ положим нормированным марковским гауссовским процессом.

Для нахождения уравнения оценки найдем функцию $F(\hat{b}, t)$:

$$F(\hat{b}, t) = \kappa z(t) s(\hat{b}, t) = \kappa u_0 z(t) \cos[\omega_0 t + \hat{\theta}(t)]. \qquad (6.124)$$

Запишем уравнение оценки, полагая вначале, что флуктуации начальной фазы отсутствуют: $\varphi(t) = \varphi_0$. Подставляя (6.123) в (6.122), получим

$$\frac{d\hat{b}(t)}{dt} = -\alpha \hat{b}(t) - \kappa z(t) u_0 \sin[\omega_0 t + \hat{\theta}(t)], \qquad (6.125)$$

где $\hat{\theta}(t) = \beta\hat{B}(t) + \varphi_0$.

Построим функциональную схему фильтра, выделяющего из реализации входного колебания $z(t)$ оптимальную оценку $\hat{B}_1(t)$. Для этого обозначим

$$\frac{\kappa}{\alpha} z(t) u_0 \sin[\omega_0 t + \hat{\theta}(t)] = y(t) \qquad (6.126)$$

и перепишем (6.125) в таком виде:

$$y(t) = \hat{B}(t) + \frac{1}{\alpha}\frac{d\hat{B}}{dt}. \qquad (6.127)$$

Легко убедиться, что если подать напряженне $y(t)$ на интегрирующую цепь RC (см., например, рис. 3.8), где $RC = 1/\alpha$, то напряженне на конденсаторе будет равно $\hat{B}(t)$.

Для того чтобы сформировать напряженне $y(t)$, достаточно иметь автогенератор на частоту ω_0, модулируемый по фазе сигналом $\hat{B}(t)$ с индексом модуляции β, и перемножитель, осуществляющий умноженне напряження автогенератора на входное колебанне $Z(t)$. Амплитуда автогенератора равна $\kappa u_0/\alpha$.

Что касается коэффициента κ, то он определяется путем решения нелинейного дифференциального уравнения, аналогичного уравненню Риккати (6.32). В данном случае для установившегося режима

$$\kappa = \frac{2\alpha}{\beta u_0}\left(\sqrt{\frac{2u_0^2\beta^2}{\alpha N_0}+1}-1\right).$$

Таким образом, получается один из возможных вариантов схемы для получения оптимальной оценки $\hat{B}(t)$, изображенный на рис. 6.15. Здесь ПГ —

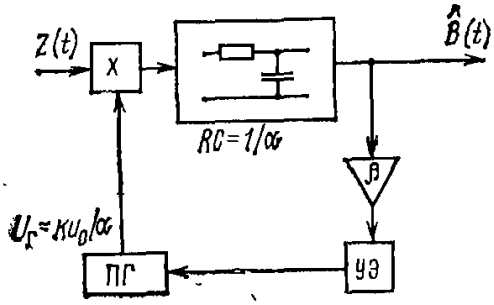


Рис. 6.15. Структурная схема оптимального приемника ФМ сигнала с постоянной фазой

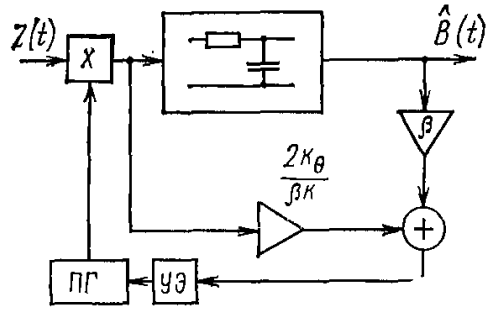


Рис. 6.16. Структурная схема оптимального приемника ФМ сигнала с блуждающей фазой

подстраиваемый генератор, фаза которого модулируется с помощью управляющего элемента УЭ. Это не что иное, как схема автоподстройки фазы автогенератора по входному колебанию $Z(t)$. Благодаря этому она обладает в известной степени свойствами саморегулирования, в частности, она мало чувствительна к точности установки начальной фазы φ_0 автогенератора и его амплитуды. Однако если начальная фаза φ подвержена значительным (по величине и скорости) флуктуациям, схема рис. 6.15 становится далеко не оптимальной.

Для того чтобы исследовать случай с «блуждающей» фазой $\varphi(t)$ подчиняющейся уравненню (6.117), учтем, что теперь в (6.125) $\hat{\theta}(t)=\beta\hat{B}+\hat{\varphi}(t)$. Теперь можно записать уравнение (6.122) для совместной оценки $\hat{B}(t)$ и $\theta(t)$:

$$\left.\begin{aligned} d\hat{B}/dt &= -\alpha\hat{B}(t) - \kappa Z(t)\,u_0 \sin[\omega_0 t + \hat{\theta}(t)];\\ d\hat{\theta}/dt &= -\alpha\beta\hat{B}(t) - \frac{2}{N_0}\kappa_\theta Z(t)\,u_0 \sin[\omega_0 t + \hat{\theta}(t)], \end{aligned}\right\} \qquad (6.128)$$

где κ_θ — коэффициент, пропорциональный апостериорной дисперсии оценки фазы, определяемый дифференциальным уравнением, которое здесь не приводится. Отметим лишь, что он зависит от спектральной плотности белого шума $W(t)$, входящей в определение процесса $\varphi(t)$ (6.117), а также от спектральной плотности аддитивного шума $N(t)$.

В этом случае в схеме фильтра на фазовый модулятор поступает помимо оценки сообщения $\hat{B}(t)$, также оценка флуктуирующей начальной фазы $\hat{\varphi}(t)$ (рис. 6.16).

При частотной модуляции в (6.123) $\theta(t)=\beta\int_0^t B(t)d(t)$.

В остальном характер всех соотношений сохраняется. Схемы оптимального нелинейного фильтра по структуре такие же, как и для фазовой модуляции, — рис. 6.15 (при постоянной начальной фазе) и рис. 6.16 (при флуктуирующей начальной фазе). Отличие заключается в том, что в подстраиваемом генераторе модулируется не фаза, а частота. Несколько отличаются также коэффициенты усиления. Эти фильтры представляют собой схемы фазовой автоподстройки частоты (ФАПЧ). Заметим, что они аналогичны схеме рис. 6.9.

Аналогично определяются схемы оптимальной нелинейной фильтрации для других видов модуляции и для более сложных каналов. Так, например, для ЧМ в канале с рэлеевскими замираниями синтезирован фильтр, основой которого является схема рис. 6.16, но с добавлением цепи автоматической регулировки усиления (АРУ), работающей так, что осуществляются подавление «слабого» сигнала и большое усиление сильного. Расчет показывает, что средняя ошибка фильтрации при этом оказывается такой же, как и в отсутствие замираний.

ГЛАВА СЕДЬМАЯ

ЦИФРОВЫЕ МЕТОДЫ ПЕРЕДАЧИ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

7.1. ОБЩИЕ СВЕДЕНИЯ О ЦИФРОВОЙ ПЕРЕДАЧЕ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Для передачи непрерывных сообщений можно воспользоваться дискретным каналом. Для этого необходимо преобразовать непрерывное сообщение в *дискретный (цифровой)* сигнал, т.е. в последовательность символов, сохранив содержащуюся в сообщении существенную часть информации, определяемую его эпсилон-энтропией. Типичными примерами цифровых систем передачи непрерывных сообщений являются системы с импульсно-кодовой модуляцией (ИКМ) и дельта-модуляцией (ДМ).

Как отмечалось в § 1.7, для преобразования непрерывного сообщения в дискретную (цифровую) форму используются операции дискретизации и квантования. Полученная таким образом последовательность квантованных отсчетов кодируется и передается по дискретному каналу, как всякое дискретное сообщение. На приемной стороне после декодирования восстанавливается (с той или иной точностью) непрерывное сообщение.

Основное техническое преимущество цифровых систем передачи перед системами непрерывного типа состоит в их высокой помехоустойчивости. Это преимущество наиболее сильно проявляется в системах передачи с *многократной ретрансляцией* (переприемом) сигналов. Типичными системами подобного типа являются кабельные и радиорелейные линии большой протяженности. В них сигналы передаются по цепи ретрансляторов, расположенных на таких расстояниях друг от друга, которые обеспечивают надежную связь. В таких системах помехи и искажения, возникающие в отдельных звеньях, как правило, накапливаются. Допустим, для простоты, что сигнал в каждом ретрансляторе только усиливается. Тогда, если аддитивные помехи в каждом звене

статистически независимы, их мощность на входе последнего звена равна сумме мощностей помех всех звеньев.

Если система состоит из k одинаковых звеньев, для обеспечения заданной верности связи необходимо обеспечить на входе каждого ретранслятора отношение сигнал/помеха в k раз больше, чем при передаче без ретрансляций. В реальных системах число ретрансляций k может достигать нескольких десятков, а иногда и сотен. В этих случаях накопление помех вдоль тракта передачи становится основным фактором, ограничивающим протяженность линии связи.

При цифровых системах передачи с целью ослабления эффекта накопления помех при передаче с ретрансляциями наряду с усилением используется *регенерация* импульсов, т. е. демодуляция с восстановлением переданных кодовых символов и повторная модуляция на переприемном пункте. При использовании регенерации аддитивная помеха со входа ретранслятора не поступает на его выход. Правда, она вызывает ошибки при демодуляции. Ошибочно принятые в одном регенераторе символы в таком виде передаются и на следующие регенераторы, так что ошибки все же накапливаются. При использовании k регенераторов и допустимой вероятности ошибки $p \ll 1$ необходимо в демодуляторе каждого регенератора обеспечить вероятность ошибки p_1 не более чем p/k. Однако, в отличие от предыдущего случая, для этого вовсе не нужно увеличивать отношение сигнал/помеха на входе регенератора в k раз. Так, например, при передаче двоичных символов в канале с аддитивным гауссовским шумом и некогерентном приеме двоичных ортогональных сигналов ошибки определяются формулой (4.88). Если требуется обеспечить $p < 10^{-5}$, то в системе без переприема для этого нужно иметь $h^2 = -2\ln(2p) = 21{,}64$. При использовании $k = 1000$ регенераторов нужно обеспечить на каждом из них $p_1 < 10^{-8}$, или $h^2 = 2\ln(2p)^{-1} \approx \approx 35{,}45$. Для этого достаточно поднять мощность сигнала в 1,64 раза, а не в 1000 раз.

При цифровой системе передачи непрерывных сообщений можно, кроме того, повысить верность применением помехоустойчивого кодирования. Высокая помехоустойчивость цифровых систем передачи позволяет осуществить практически неограниченную по дальности связь при использовании каналов сравнительно невысокого качества.

Другим существенным преимуществом цифровых систем передачи информации является широкое использование в аппаратуре преобразования сигналов современной элементной базы цифровой вычислительной техники и микроэлектроники. Больше того, на цифровой основе могут быть объединены в единой системе сигналы передачи данных с сигналами передачи речи и телевидения. Возможность приведения всех видов передаваемой информации к цифровой форме позволит осуществить интеграцию систем передачи и систем коммутации. Простота сочленения цифрового канала с ЭВМ позволяет существенно расширить область исполь-

зования вычислительной техники при построении аппаратуры связи и автоматизации управления сетями связи и ускорить тем самым решение проблемы построения Единой автоматизированной сети связи страны.

Отмеченные и другие преимущества цифровых систем передачи непрерывных сообщений позволяют предположить, что в бли-

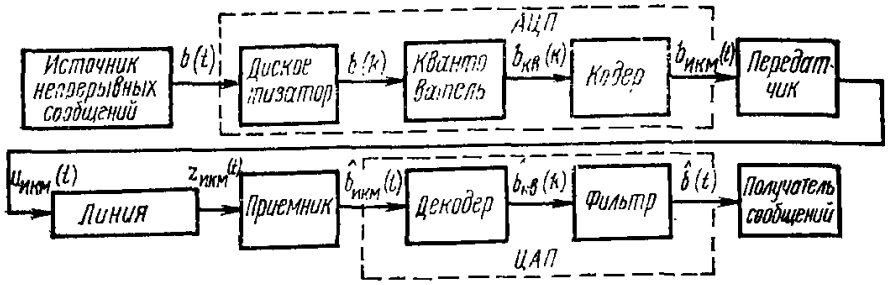


Рис. 7.1. Структурная схема системы цифровой передачи непрерывных сообщений

жайшие годы эти системы найдут не менее широкое применение, чем непрерывные (аналоговые) системы связи.

Перейдем теперь к рассмотрению структурной схемы цифрового канала передачи непрерывных сообщений (рис. 7.1). В отличие от непрерывного канала передачи, в составе цифрового канала предусмотрены устройства для преобразования непрерывного сообщения в цифровую форму — аналого-цифровой преобразователь (АЦП) на передающей стороне и устройства преобразования цифрового сигнала в непрерывный — цифро-аналоговый преобразователь (ЦАП) на приемной стороне.

Преобразование аналог — цифра состоит из трех операций (рис. 7.2): сначала непрерывное сообщение подвергается дискретизации по времени

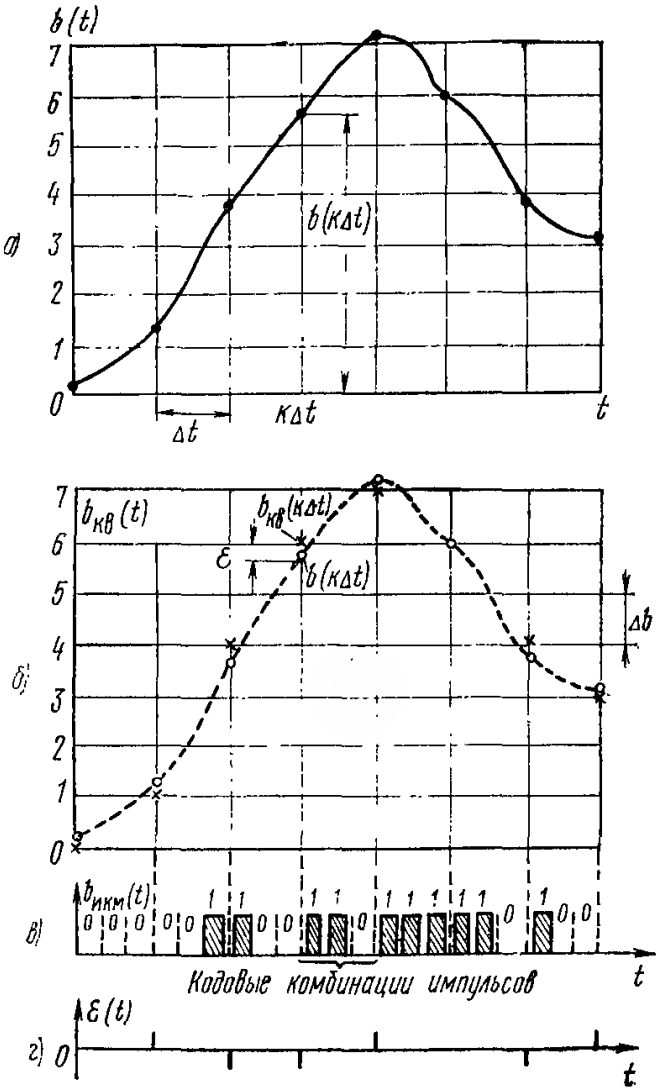


Рис. 7.2. Преобразование непрерывного сообщения в последовательность двоичных импульсов

через интервалы Δt (рис. 7.2*а*); полученные отсчеты мгновенных значений $b(k\Delta t)$ квантуются (рис. 7.2*б*); наконец, полученная последовательность квантованных значений $b_{кв}(k\Delta t)$ передаваемого сообщения представляется посредством кодирования в виде последовательности m-ичных кодовых комбинаций (рис. 7.2*в*). Такое преобразование называется импульсно-кодовой модуляцией (ИКМ). Чаще всего кодирование здесь сводится к записи номера уровня в двоичной системе счисления. В дальнейшем будем рассматривать цифровые системы, в которых непрерывное сообщение преобразовано в последовательность кодовых комбинаций, составленных из двоичных символов.

Полученный с выхода АЦП сигнал ИКМ поступает или непосредственно в линию связи или на вход передатчика (модулятора), где последовательность двоичных импульсов преобразуется в радиоимпульсы.

На приемной стороне линии связи последовательность импульсов после демодуляции и регенерации в приемнике поступает на цифро-аналоговый преобразователь ЦАП, назначение которого состоит в обратном преобразовании (восстановлении) непрерывного сообщения по принятой последовательности кодовых комбинаций. В состав ЦАП входят декодирующее устройство, предназначенное для преобразования кодовых комбинаций в квантованную последовательность отсчетов, и сглаживающий фильтр, восстанавливающий непрерывное сообщение по квантованным значениям.

Преобразование непрерывных сообщений в цифровую форму в системах ИКМ, как уже отмечалось, сопровождается *округлением* мгновенных значений до ближайших разрешенных уровней квантования. Возникающая при этом погрешность представления является неустранимой, но *контролируемой* (так как не превышает половины шага квантования). Выбрав достаточно малый шаг квантования, можно обеспечить эквивалентность по заданному ε-критерию (см. § 6.1) исходного и квантованного сообщений. Погрешность (ошибку) квантования, представляющую собой разность между исходным сообщением и сообщением, восстановленным по квантованным отсчетам, называют *шумом квантования.*

7.2. ПОМЕХОУСТОЙЧИВОСТЬ ИМПУЛЬСНО-КОДОВОЙ МОДУЛЯЦИИ

Одной из причин, приводящих к отличию принятого сообщения от переданного в системе с ИКМ, является шум квантования, другой — помехи в канале, которые накладываются на передаваемые символы кодовых комбинаций и могут вызвать ошибки. Ошибки в символах (при отсутствии избыточности) приводят к ошибочному декодированию всей кодовой комбинации.

В результате ошибочного декодирования символа действительно переданное дискретное значение сообщения заменяется другим возможным (не обязательно ближайшим); погрешность зависит от того, какие из символов кодовой комбинации приняты

с ошибкой. Назовем эту составляющую шума шумом ложных импульсов. Таким образом, при оценке помехоустойчивости необходимо учитывать суммарный шум как за счет квантования, так и за счет ложных импульсов при декодировании.

Шум квантования не связан с помехами в канале и целиком определяется выбором числа уровней квантования. Его можно сделать сколь угодно малым, увеличивая число уровней. При этом придется увеличивать число кодовых символов, приходящихся на каждый отсчет, а следовательно, сокращать длительность символа и расширять спектр сигнала в канале. Таким образом, также, как и при помехоустойчивых аналоговых видах модуляции, снижение этого шума достигается за счет расширения спектра сигнала.

Шум ложных импульсов является аномальным (см. § 6.4). Он целиком определяется помехами в канале и видом модуляции несущей. При расширении спектра сигнала мощность аномального шума, как правило, возрастает.

Мощность шума квантования. Для определения мощности шума квантования представим реализацию $b(t)$ непрерывного сообщения ее разложением в ряд Котельникова:

$$b(t)=\sum_{k=-\infty}^{\infty} b(k\Delta t)\,\frac{\sin 2\pi F_c(t-k\Delta t)}{2\pi F_c(t-k\Delta t)}. \tag{7.1}$$

После фильтрации квантованных отсчетов получим функцию $b_{\text{кв}}(t)$, приближенно отображающую исходное сообщение $b(t)$:

$$b_{\text{кв}}(t)=\sum_{k=-\infty}^{\infty} b_{\text{кв}}(k\Delta t)\,\frac{\sin 2\pi F_c(t-k\Delta t)}{2\pi F_c(t-k\Delta t)}. \tag{7.2}$$

Представим, далее, квантованное значение отсчета $b_{\text{кв}}(k\Delta t)$ в виде суммы

$$b_{\text{кв}}(k\Delta t)=b(k\Delta t)+\varkappa_k\Delta b, \tag{7.3}$$

где $\varkappa_k$ — безразмерная случайная величина, лежащая в интервале $-0{,}5<\varkappa_k<0{,}5$. В расчетах часто принимают, что погрешность квантования имеет равномерное распределение (при большом числе уровней квантования это близко к истине):

$$\begin{cases} w(\varkappa)=1 & \text{при } |\varkappa|<0{,}5; \\ w(\varkappa)=0 & \text{при } |\varkappa|>0{,}5. \end{cases} \tag{7.4}$$

Обозначив отсчетную функцию Котельникова в (7.1) через $\psi_k(t)$, перепишем (7.2) в виде

$$b_{\text{кв}}(t)=\sum_{k=-\infty}^{\infty}[b(k\Delta t)+\varkappa_k\Delta b]\,\psi_k(t),$$

или

$$b_{\text{кв}}(t)=b(t)+\Delta b\sum_{k=-\infty}^{\infty}\varkappa_k\,\psi_k(t). \tag{7.5}$$

Обозначив погрешность квантования (шум квантования) через

$$\varepsilon(t) = \Delta b \sum_{k=-\infty}^{\infty} \varkappa_k \psi_k(t), \tag{7.6}$$

для (7.5) имеем

$$b_{\text{кв}}(t) = b(t) + \varepsilon(t). \tag{7.7}$$

Определим теперь среднюю мощность шума квантования.

$$\overline{\varepsilon^2(t)} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \overline{\varepsilon^2(t)}\, dt =$$

$$= (\Delta b)^2 \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \overline{\sum_{k=-T/(2\Delta t)}^{T/(2\Delta t)} \varkappa_k \psi_k(t) \sum_{l=-T/(2\Delta t)}^{T/(2\Delta t)} \varkappa_l \psi_l(t)}\, dt =$$

$$= (\Delta b)^2 \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \sum_{k=-T/(2\Delta t)}^{T/(2\Delta t)} \sum_{l=-T/(2\Delta t)}^{T/(2\Delta t)} \overline{\varkappa_k \varkappa_l}\, \psi_k(t)\, \psi_l(t)\, dt =$$

$$= (\Delta b)^2 \lim_{T\to\infty} \frac{1}{T} \sum_{k=-T/(2\Delta t)}^{T/(2\Delta t)} \sum_{l=-T/(2\Delta t)}^{T/(2\Delta t)} \overline{\varkappa_k \varkappa_l} \int_{-\infty}^{\infty} \psi_k(t)\, \psi_l(t)\, dt. \tag{7.8}$$

Принимая во внимание, что

$$\int_{-\infty}^{\infty} \psi_k(t)\, \psi_l(t)\, dt = \begin{cases} 1/(2F_c) = \Delta t & \text{при } l = k \\ 0 & \text{при } l \neq k, \end{cases}$$

получим

$$\overline{\varepsilon^2(t)} = (\Delta b)^2 \lim_{T\to\infty} \frac{1}{T} \sum_{k=-T/(2\Delta t)}^{T/(2\Delta t)} \overline{\varkappa_k^2}\, \Delta t = \lim_{T\to\infty} (\Delta b)^2 \frac{\Delta t}{T} \frac{T}{\Delta t} \overline{\varkappa_k^2} = (\Delta b)^2 \overline{\varkappa_k^2}. \tag{7.9}$$

Из (7.4) имеем

$$\overline{\varkappa_k^2} = \mathrm{D}\{\varkappa\} = \int_{-0,5}^{0,5} \varkappa^2\, d\varkappa = \left.\frac{\varkappa^3}{3}\right|_{-0,5}^{0,5} = 1/12. \tag{7.10}$$

Следовательно, средняя мощность шума квантования

$$\overline{\varepsilon^2(t)} = (\Delta b)^2 \mathrm{D}(\varkappa_k) = (\Delta b)^2/12, \tag{7.11}$$

а отношение средних мощностей сообщения и шума квантования

$$P_B/P_\varepsilon = \overline{B^2(t)}/\overline{\varepsilon^2(t)} = 12\,\overline{B^2(t)}/(\Delta b)^2. \tag{7.12}$$

Выразим Δb через число уровней L. Полагая, что $B(t)$ — нормированное сообщение $-1 < B(t) < 1$ (см. § 6.5), получим

$$\Delta b = (B_{\text{макс}} - B_{\text{мин}})/(L-1) = 2/(L-1). \tag{7.13}$$

С другой стороны, согласно (6.79) $\overline{B^2(t)} = P_B = 1/\Pi^2$, где Π — пик-фактор сообщения. Тогда

$$\frac{P_B}{P_\varepsilon} = \frac{\overline{B^2(t)}}{\varepsilon^2(t)} = \frac{1}{\Pi^2}\,\frac{12}{(\Delta b)^2} = \frac{3(L-1)^2}{\Pi^2} = \frac{3(2^n-1)^2}{\Pi^2}, \qquad (7.14)$$

где n — число символов на отсчет при безызбыточном двоичном коде. Из (7.14) следует, что верность квантованного сообщения зависит от числа уровней квантования. Выбирая его достаточно большим, можно уменьшить относительное значение шума квантования до любой допустимой величины. В табл. 7.1 приведена

Таблица 7.1

Число уровней квантования L	8	16	32	64	128	256	512	1024	2048
Количество символов в кодовой комбинации n	3	4	5	6	7	8	9	10	11
Относительная мощность шума квантования $-20\lg(L-1)$	—16,9	—23,5	—29,8	—36,0	—42,1	—48,1	—54,2	—60,2	—66,2

зависимость отношения P_B/P_ε от числа уровней квантования при равномерном распределении $w(B) = \frac{1}{2}$ $(-1 < B < 1)$, когда $\Pi = \sqrt{3}$.

Из табл. 7.1 следует, что добавление каждого двоичного символа в кодовой комбинации (увеличение разрядности кода) улучшает отношение P_B/P_ε приблизительно на 6 дБ. С другой стороны, увеличение разрядности требует повышения быстродействия многоразрядных кодирующих устройств, а также соответствующего расширения полосы частот канала передачи.

Чаще всего распределение сообщения не является равномерным и в полученные здесь цифры необходимо внести поправку. Для этого данные таблицы нужно уменьшить по модулю на $20\lg(\Pi/\sqrt{3})$ дБ. При $\Pi = 3$ (телефонные сообщения) это составляет 4,8 дБ, а при $\Pi = 10$ (симфоническая музыка) 15,2 дБ.

Важной особенностью шума квантования, отличающей его от аддитивных шумов, является то, что он возникает одновременно с появлением сообщения $B(t)$. С технической точки зрения шум квантования можно трактовать как разновидность нелинейных искажений, возникающих в процессе квантования. Поэтому шум квантования не изменяется при ретрансляции сигналов, т. е. не накапливается. Его непосредственное измерение затруднительно, и потому для количественной оценки используют расчетные приемы, один из которых рассмотрен выше.

Воздействие шума квантования на принимаемые сообщения можно заметно уменьшить, применяя неравномерное квантование, при котором большие уровни сообщения квантуются с бо́льшим шагом, а низкие уровни — с меньшим шагом (рис. 7.3). Очевидно, что шум квантования при этом коррелирован с сообщением и имеет тем меньшую мгновенную мощность, чем меньше уровень сообщения. Это позволяет при том же числе уровней лучше различать слабые отрезки сообщения. Если распределение вероятностей сообщения такое, что большие уровни встречаются значительно реже, чем малые (что имеет место, например, при передаче речи), то неравномерное квантование сводится к тому, что часто возникающие значения $B(t)$ передаются с меньшей ошибкой квантования, а редко возникающие значения $B(t)$ — с большей ошибкой квантования. В результате усреднения по всем значениям дисперсия ошибки квантования будет уменьшена.

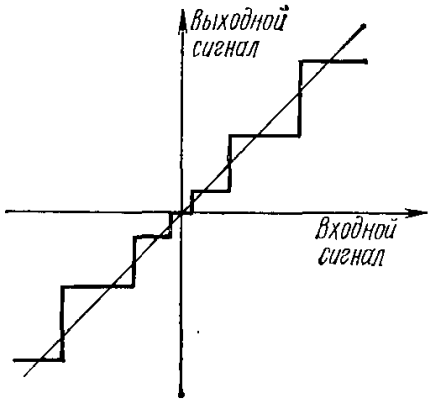


Рис. 7.3. Характеристика квантователя с неравномерным шагом

Одним из распространенных способов неравномерного квантования является квантование с *компандированием* сигнала. Компандерная система представляет собой комплекс из двух нелинейных преобразователей с взаимно обратными характеристиками — компрессора и экспандера (рис. 7.4). Динамический диапа-

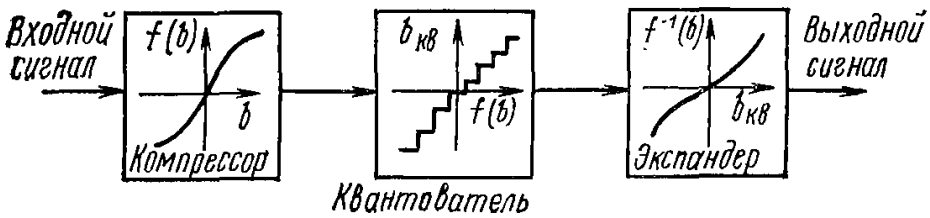


Рис. 7.4. Схема, поясняющая принцип компандирования

зон входного сигнала «сжимается» с помощью *компрессора* на передающей стороне, затем сигнал равномерно квантуется. Комбинация этих операций эквивалентна неравномерному квантованию, причем закон изменения шага квантования определяется нелинейной характеристикой компрессора. После кодирования и передачи по линии связи на приемной стороне производится обратное преобразование, восстанавливаются отсчеты с равномерным шагом квантования. Затем они подвергаются экспандированию, при котором восстанавливается исходный динамический диапазон. В настоящее время более перспективным считается компандирование в цифровой форме, т. е. непосредственное неравномерное квантование.

Шум ложных импульсов при декодировании. Обозначим вероятность ошибочного приема одного символа кодовой комбинации через p. Эта вероятность зависит от вида модуляции и находится по формулам гл. 4. Предполагая ошибки при приеме символов независимыми, запишем выражение для $P(\Delta)$ — вероятности того, что кратность сочетания ошибок составит Δ:

$$P(\Delta) = C_n^{\Delta} p^{\Delta} (1-p)^{n-\Delta}. \tag{7.15}$$

Вероятность того, что кодовая комбинация принимается хотя бы с одной ошибкой при $np \ll 1$, равна

$$[1-(1-p)^n] \approx np. \tag{7.16}$$

При декодировании каждый символ кодовой комбинации, в зависимости от занимаемого им места, дает определенный «вклад» в декодируемое сообщение. Если используется двоичный код, то ошибка в младшем разряде кодовой комбинации вызывает погрешность в выходном сообщении, равную шагу квантования Δb; ошибка во втором символе приводит к появлению в выходном сообщении погрешности $2\Delta b$ и т. д. При этом средняя мощность шума, обусловленного действием ложных импульсов, равна

$$\overline{\varepsilon_{\text{ли}}^2} = [1-(1-p)^n] \frac{(\Delta b)^2}{n} \sum_{i=1}^{n} 2^{2(i-1)} \approx p(\Delta b)^2 \sum_{i=1}^{n} 2^{2(i-1)}. \tag{7.17}$$

При фиксированном значении $n = \log L$ шум ложных импульсов зависит только от вероятности ошибок p, которая, в свою очередь, определяется отношением мощностей сигнала и помехи в канале и видом модуляции.

Как уже отмечалось, в отличие от шума квантования, шум ложных импульсов накапливается при ретрансляции. Однако в правильно спроектированных системах с ИКМ мощность сигнала превышает пороговую, при которой аномальным шумом ложных импульсов, по сравнению с шумом квантования, можно пренебречь. При этом условии верность приема практически определяется шумом квантования и может быть сколь угодно большой, если число уровней достаточно велико.

Очевидно, что в системе с ИКМ так же, как и в других помехоустойчивых системах передачи непрерывных сообщений, имеет место порог помехоустойчивости, т. е. верность приема резко ухудшается, если мощность сигнала упадет ниже пороговой. Из сказанного выше ясно, что эта пороговая мощность увеличивается с ростом числа ретрансляторов, впрочем очень медленно.

Легко также сообразить, что пороговая мощность увеличивается и с ростом числа уровней квантования. Основная причина этого заключается в том, что чем больше число уровней, тем больше кодовых символов должно приходиться на один отсчет и следовательно, тем меньше длительность передачи одного симво-

ла. Поскольку же вероятность ошибки определяется энергией элемента сигнала, то при сокращении его длительности приходится увеличивать его мощность. Впрочем, это увеличение пороговой мощности также невелико по сравнению с соответствующим уменьшением шума квантования. Так, с переходом от 128 уровней кваитования к 256 шум квантования уменьшается на 6 дБ (см. табл. 7.1). При этом вместо семи символов в кодовой комбинации приходится передавать восемь, так что длительность импульса уменьшится в $8/7 \approx 1{,}14$ раза. Для того чтобы сохранить прежнюю вероятность ошибки, нужно увеличить мощность сигнала тоже в 1,14 раза, т. е. всего лишь на 0,6 дБ.

Следует также отметить, что слабый шум ложных импульсов, имеющий место при работе над порогом помехоустойчивости, воспринимается (в телефонных системах) как более или меиее редкие отдельные щелчки. Если мощность сигнала упадет и окажется ниже порога, эти щелчки становятся частыми и сливаются в сплошной шумовой фон. Аналогичная картина имеет место и для аномального шума в аналоговых системах (например, ЧМ).

Высокая помехоустойчивость ИКМ систем достигается за счет расширения спектра ИКМ сигнала по сравнению со спектром исходиого сообщения. Найдем вначале минимальную ширину спектра ИКМ сигнала при основании кода $m=2$. Если ширина спектра исходного сообщения равна F_c, то минимальная частота дискретизации, в соответствии с теоремой Котельникова, равна $2F_c$. Каждый отсчет после квантования может принимать $L=2B_{макс}/\Delta b+1$ возможных дискретных значений и заменяется при кодировании комбинацией из $n=\log L$ двоичных импульсов. Следовательно, длительность каждого импульса не может быть больше, чем $\tau_и=1/[2F_c \log L]$, а необходимая полоса частот определяется как $F' \approx 1/(2\tau_и)=F_c \log L$. При двухполосной АМ сигнал ИКМ-АМ будет занимать полосу частот

$$F = 2F' = 2F_c \log L. \tag{7.18}$$

Поскольку при ИКМ верность передачи определяется числом уровней квантования, то увеличение вериости сопровождается расширеиием спектра ИКМ сигнала по логарифмическому закону. Так, увеличение L в 2 раза приводит к увеличению ширины спектра сигнала в $\log 2L/\log L=(1+1/\log L)$ раз. Ширина спектра ИКМ сигнала зависит от основания кода m: при $m=2$ ширина спектра ИКМ сигнала наибольшая; при увеличении m ширина спектра уменьшается.

Из изложенного следует, что в системе передачи с ИКМ, как и в помехоустойчивых аналоговых системах модуляции, производится «обмен» мощиости сигнала на полосу частот. Принципиальная возможность такого обмена была показана в гл. 1. Однако в системе с ИКМ этот обмеи осуществляется значительно эффективнее, чем в системах с аналоговой модуляцией. Действительно, как было показано на ряде примеров в гл. 6, в помехоустойчивых системах модуляции, таких, как ЧМ, ФМ, ВИМ, отно-

шение мощности сообщения к мощности шума на выходе растет пропорционально квадрату ширины спектра сигнала (если мощность сигнала выше пороговой). В системе с ИКМ имеет место значительно более быстрый, экспоненциальный рост этого отношения. Действительно, ширина спектра пропорциональна числу n символов в кодовой комбинации, тогда как мощность шума квантования уменьшается в соответствии с (7.13) почти пропорционально 2^{2n}. Другими словами, пропорционально ширине спектра растет выигрыш системы, выраженный в децибелах.

Как было показано в § 6.8, такой же характер зависимости верности от ширины спектра должен быть в идеальной системе модуляции, так что в этом отношении ИКМ ведет себя как идеальная система. Более подробный анализ приводит к выводу, что при одинаковой ширине спектра выигрыш в ИКМ приблизительно на 8 дБ меньше, чем в теоретически идеальной системе. В настоящее время не существует систем модуляции, более близких к идеальной, если спектр передаваемого сообщения равномерный. Поэтому система с ИКМ широко используется в тех случаях, когда высокую верность необходимо обеспечить с минимальной затратой мощности передатчика, например, в спутниковых системах. Более полное сравнение различных систем связи по их эффективности будет дано в § 7.4.

7.3. КОДИРОВАНИЕ С ПРЕДСКАЗАНИЕМ

В ряде практически важных случаев, например, при передаче речи, телевизионных изображений и т. п., между отсчетами передаваемых сообщений имеются статистические, в частности, корреляционные взаимосвязи. Наличие таких взаимосвязей позволяет повысить эффективность систем передачи информации[1]. Рассмотрим один из распространенных способов передачи — спо-

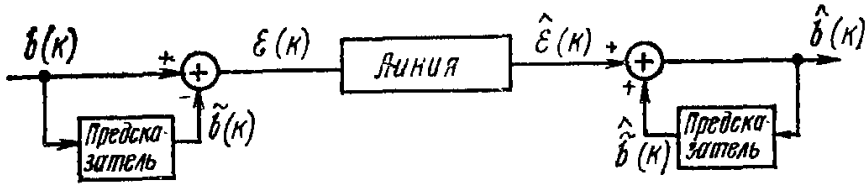


Рис. 7.5. Структурная схема системы связи с предсказанием

соб передачи с предсказанием (рис. 7.5). Последовательность коррелированных отсчетов $b(k\Delta t)=b(k)$ исходного сигнала подается на один из входов вычитающего устройства, а на его дру-

[1] Как было показано в § 2.7, значения отсчетов, взятых через интервал Котельникова, взаимно некоррелированны, если спектр сообщения в занимаемой им полосе частот равномерен. Но на практике приходится часто передавать сообщения с неравномерным спектром (например, речь, телевидение), да к тому же частота отсчетов по ряду соображений выбирается несколько выше котельниковской. Поэтому корреляция между отсчетами обычно не равна нулю.

гой вход поступает сигнал предсказания $\widetilde{b(k)}$, сформированный из предыдущих отсчетов. Полученный таким образом сигнал ошибки предсказания $\varepsilon(k)$ поступает в тракт передачи. Поскольку в сигнале ошибки как раз и содержатся новые сведения, представляющие разность между истинным и предсказанным значениями, то такой способ передачи называется передачей с предсказанием. Для упрощения считаем, что помехи в канале ошибок не создают. На приемном конце имеется такой же предсказатель, как и на передающем. Поскольку он оперирует с теми же предыдущими отсчетами, предсказанное им значение нового отсчета $\hat{b}(k)$ будет таким же, как и на передатчике. Добавив к нему принятое значение ошибки предсказания $\varepsilon(k)$, можно восстановить истинный отсчет:

$$b(k)=\hat{b}(k)+\varepsilon(k). \tag{7.19}$$

Очевидно, что чем сильнее корреляционные связи, тем точнее можно сформировать сигнал предсказания и, следовательно, тем меньшая энергия потребуется для передачи сигнала ошибки по сравнению с исходным сигналом. При сильных корреляционных связях, а также при марковских сообщениях значение $\hat{b}(k)$ может быть предсказано по одному предыдущему отсчету:

$$\hat{b}(k)=f[b(k-1)]. \tag{7.20}$$

Обычно можно получить достаточно малую ошибку предсказания, положив просто

$$\hat{b}(k)=b(k-1), \tag{7.21}$$

откуда

$$\varepsilon(k)=b(k)-b(k-1). \tag{7.22}$$

Действительно, для средней энергии сигнала ошибки по ансамблю имеем

$$E_\varepsilon=\overline{[B(k)-B(k-1)]^2}=\overline{B^2(k)}+\overline{B^2(k-1)}-\overline{2B(k)\,B(k-1)}=$$
$$=2\,\overline{B^2(k)}-\overline{2\,B(k)\,B(k-1)}=2\,E_B(1-R), \tag{7.23}$$

где $R=[\overline{B(k)B(k-1)}]/E_B$ — коэффициент корреляции между отсчетами. Из (7.23) следует, что при $R>0{,}5$ энергия сигнала ошибки E_ε меньше энергии исходного сигнала E_B.

При цифровых системах передачи отсчеты сигнала ошибки подвергаются обычным операциям квантования и кодирования. Результатом такого преобразования на передающей стороне является импульсно-кодовое представление сигнала ошибки, или сигнал *дифференциальной импульсно-кодовой модуляции* (ДИКМ). В настоящее время известно большое число вариантов технического осуществления кодирования с предсказанием. Основное их различие сводится к различию операций формирования сигнала ошибки: в одних системах сигнал ошибки формируется в аналоговой форме, а затем кодируется, в других — вначa-

ле кодируется исходный аналоговый сигнал, а затем формируется сигнал ошибки.

В системах с ДИКМ обычно применяют неравномерное квантование сигнала ошибки, так как наиболее вероятны малые ошибки. Поскольку мощность шума квантования составляет определенную долю мощности квантуемого процесса, а мощность ошибки предсказания, как правило, существенно меньше мощности сообщения, шум квантования при ДИКМ меньше, чем при обычной ИКМ при том же числе уровней.

Как показывает анализ, система ДИКМ обеспечивают одинаковое с ИКМ качество передачи при меньшем числе символов в кодовой комбинации (на один — два и более символов в зависимости от степени корреляции отсчетов).

Шум ложных импульсов также имеет место при ДИКМ, причем он ухудшает верность приема в большей мере, чем при обычной ИКМ. Дело в том, что ошибочный прием кодовой комбинации ведет к ошибочному приему не только одного отсчета сообщения, но и ряда последующих отсчетов, поскольку предсказанные значения на приемной стороне будут отличаться от предсказанных на передатчике. Поэтому допустимая вероятность ошибок при ДИКМ меньше, чем при ИКМ, и, следовательно, пороговая мощность сигнала выше. Существуют различные способы уменьшения влияния «размножения» ошибок.

Корреляция между отсчетами возрастает по мере сокращения интервала между ними. Поэтому при большой частоте дискретизации число уровней квантования сигнала ошибки можно уменьшить до двух и перейти к одноразрядным системам. Такой способ кодирования называют *дельта-модуляцией.*

Квантованный сигнал ошибки имеет вид

$$\varepsilon_{\text{кв}}(k) = \gamma(k)\,\Delta b, \tag{7.24}$$

где

$$\gamma(k) = \begin{cases} +1, & \text{если} \quad \varepsilon(k) \geqslant 0; \\ -1, & \text{если} \quad \varepsilon(k) < 0. \end{cases}$$

По существу, сигнал на выходе дельта-модулятора содержит лишь сведения о полярности (знаке) сигнала ошибки. На приемной стороне интегратор прибавляет или вычитает Δb, обеспечивая, таким образом, уменьшение погрешности между исходными значениями отсчетов и восстановленными значениями. Принцип формирования сигнала дельта-модуляции показан на рис. 7.6.

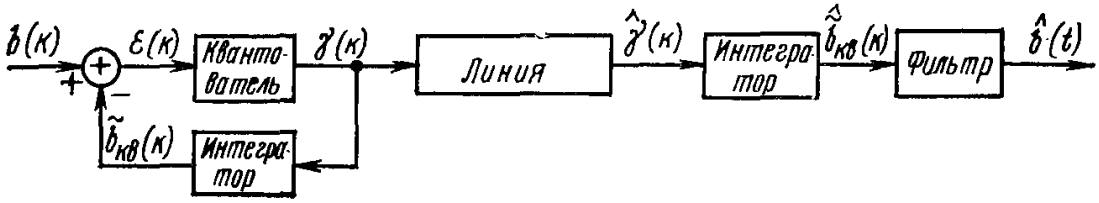


Рис. 7.6. Структурная схема системы с дельта-модуляцией

Отсчеты $b(k)$ передаваемого сообщения сравниваются с квантованным отсчетом $b_{\text{кв}}(k-1)$, полученным в результате суммирования всех предыдущих квантованных сигналов ошибки:

$$b_{\text{кв}}(k-1) = \sum_{i=0}^{n-1} \varepsilon_{\text{кв}}(i) = \Delta b \sum_{i=0}^{n-1} \gamma(i).$$

Если $b(k) > b_{кв}(k-1)$, квантователь формирует значение $\gamma(k)=1$, в противном случае $\gamma(k)=-1$. Полученные символы γ передаются по тракту связи (рис. 7.7*б*). Одновременно импульсы $\Delta b\gamma(i)$ поступают на сумматор (интегратор) для формирования квантованного отсчета $b_{кв}(i)$, который, как указывалось выше, сравнивается с очередным отсчетом сообщения. На выходе сумматора квантованный сигнал $\tilde{b}_{кв}(t)$ имеет вид ступенчатой функции (рис. 7.7*а*). Каждый импульс $+1$ увеличивает, а каждый импульс -1 уменьшает ступенчатую функцию на один шаг квантования (рис. 7.7*а*), т. е. при ДМ соседние значения ступенчатой функции различаются обязательно на величину *одного* шага квантования.

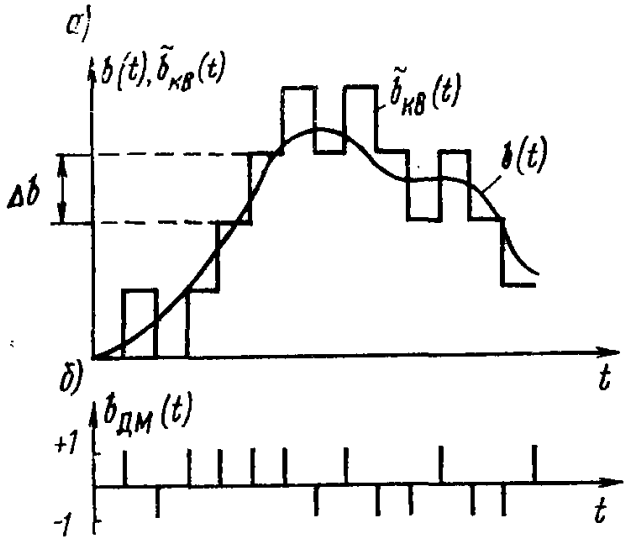


Рис. 7.7. Графики, поясняющие принцип формирования сигналов дельта-модуляции

Операцию декодирования сигнала ДМ на приемной стороне выполняет сумматор (интегратор), такой же, как и на передающей стороне, на выходе которого (при отсутствии ложных импульсов) получается ступенчатое напряжение $\tilde{b}_{кв}(t)$. После его сглаживания с помощью фильтра получим функцию $\hat{b}(t)$, достаточно близкую к $b(t)$. Разность $\hat{b}(t)-b(t)$ представляет шум квантования. Его величина тем меньше, чем выше частота дискретизации и чем меньше шаг Δb. Однако шаг квантования Δb нельзя выбирать чересчур малым. В противном случае возникают дополнительные искажения, называемые «перегрузкой по наклону», вызванные тем, что ступенчатая функция не успевает следить за быстрыми изменениями сообщения $b(t)$. Это иллюстрирует рис. 7.8. Если принять, что максимальное значение крутизны изменения сообщения составляет $|b'(t)|_{макс}$, то для неискаженной передачи необходимо выполнить условие $|b'(t)|_{макс}\Delta t \leqslant \Delta b$. Если практически максимальное значение сообщения по модулю не превышает $|b(t)|_{макс}$, а число различных уровней квантования, определяющее шум квантования, равно L, то отсюда следует

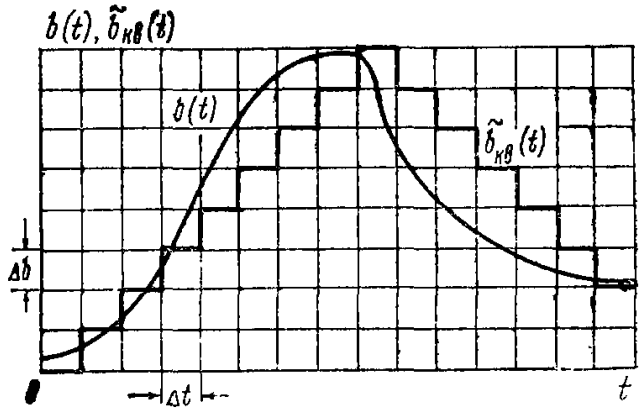


Рис. 7.8. График, поясняющий возникновение искажений из-за перегрузки по наклону при дельта-модуляции

$$\Delta t \leqslant \frac{1}{L} \frac{|b(t)|_{\text{макс}}}{|b'(t)|_{\text{макс}}}. \qquad (7.25)$$

По сравнению с ИКМ и ДИКМ сигналы ДМ имеют значительно более высокую частоту следования отсчетов. Однако на каждый отсчет при ДМ передается один импульс, а при ИКМ — несколько, в зависимости от числа уровней. Как показывает анализ, при одинаковой верности передачи частота следования импульсов при ИКМ и ДМ также примерно одинакова. Поэтому обе эти системы занимают приблизительно одинаковую полосу частот. Существенным преимуществом систем передачи с ДМ является сравнительная простота кодирующих и декодирующих устройств. Кроме того, шум ложных импульсов при ДМ меньше, чем при ИКМ (при той же вероятности ошибки в канале), поскольку при ДМ каждая ошибка изменяет уровень сигнала только на $\pm 2\Delta b$. К недостаткам ДМ можно отнести явление размножения ошибок, создаваемых ложными импульсами (такое же, как и при ДИКМ), а также возможность перегрузки по наклону.

Отметим, что в последние годы успешно разрабатываются многочисленные разновидности ДИКМ и ДМ, в частности адаптивные системы, в которых с целью уменьшения шумов квантования используется переменный шаг квантования в зависимости от текущих статистических характеристик передаваемого сообщения.

7.4. ЭФФЕКТИВНОСТЬ СИСТЕМ СВЯЗИ

Различные системы связи для передачи дискретных и непрерывных сообщений, несмотря на их различие как по назначению, так и по способам реализации, характеризуются определенными качественными показателями, которые были перечислены в § 1.7. Основными из них являются скорость и верность передачи информации.

Для обеспечения заданной скорости передачи информации R и заданной верности приходится затрачивать некоторую мощность сигнала и занимать определенную полосу частот в канале связи. Какая мощность и какая полоса частот при этом потребуются, зависит от используемой системы связи. Представляет интерес сравнить между собой различные системы связи по степени эффективности использования ими основных ресурсов канала — пропускной способности, мощности сигнала и занимаемой полосы частот. Ниже это сравнение производится для некоторых систем в простейшем случае, когда сигналы передаются в канале с аддитивным гауссовским белым шумом, а помехоустойчивое кодирование не применяется. Верность передачи будем считать заданной; она характеризуется при передаче дискретных сообщений вероятностью ошибки, а при передаче непрерывных сообщений — отношением сигнал/шум на выходе демодулятора.

Для оценки эффективности систем связи введем коэффициент использования канала по мощности (энергетическую эффективность)

$$\beta = \frac{R}{P_c/N_0} \qquad (7.26)$$

и коэффициент использования канала по полосе частот (частотную эффективность)

$$\gamma = R/F. \qquad (7.27)$$

Характеристикой эффективности систем связи является также коэффици-

еит использования канала по пропускной способности (информационная эффективиость):

$$\eta = R/C. \tag{7.28}$$

С учетом формулы Шеинона (3.64)

$$C = F \log(\rho + 1), \tag{7.29}$$

где $\rho = P_c/P_ш$ — отношение мощностей сигнала и шума в полосе F, получаем следующие выражения:

$$\eta = \frac{\gamma}{\log(\gamma/\beta + 1)}, \quad \gamma = \rho\beta. \tag{7.30}$$

Согласно теореме Шеииоиа при соответствующих способах передачи (кодироваиия и модуляции) и приема (демодуляции и декодирования) величина η может быть сколь угодно близка к единице. Прн этом ошибка может быть сколь угодно малой. В этом случае имеем предельную зависимость между β и γ:

$$\beta = \gamma/(2^{\gamma} - 1). \tag{7.31}$$

Удобно эту зависимость представить графически в виде кривой на плоскости $\beta\gamma$ (рис. 7.9). Полученная кривая является предельной и отражает наилучший обмен между β и γ. Следует заметить, что частотная эффективность γ может изменяться в широких пределах (от 0 до ∞ при аналоговой передаче и

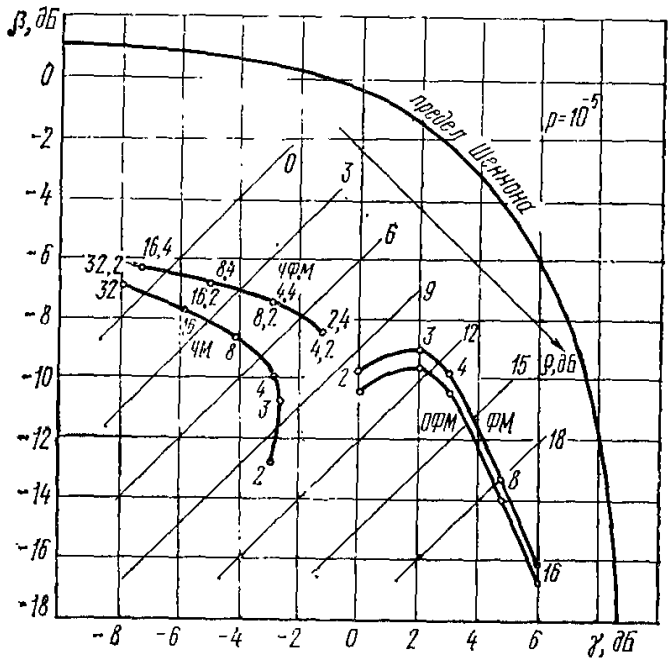


Рис. 7.9. Кривые энергетической и частотной эффективности иекоторых систем передачи дискретных сообщений

от 0 до $2 \log m$ при дискретной передаче), в то время как энергетическая эффективность β ограничена сверху: при $F \to \infty$ $\beta_{макс} = 1/2 \ln 2$. При логарифмическом масштабе, который и принят на рис. 7.9, в соответствии с соотношением $\gamma = \beta\rho$ линии одинаковых значений превышения сигнала над шумом ρ представляют собой прямые с углом наклона 45°.

В реальных системах ошибка всегда имеет конечное значение и $\eta < 1$. В этих случаях при $p = \text{const}$ можно определить отдельно β и γ и построить кривые $\beta = f(\gamma)$. В координатах β и γ каждому варианту реальиой системы будет соответствовать точка на плоскости. Все эти точки (кривые) располагаются ниже предельной кривой Шениона. Ход этих кривых завнсит от вида сигналов

(модуляции), кода и способа обработки сигналов. Цифры на кривых рис. 7.9 указывают число позиций сигнала m. Кривые рассчитаны на основании формул гл. 4 для оптимального приема сигиалов при равной априорной вероятности их передачи и вероятности ошибки $p=10^{-5}$. Если T — время, затрачиваемое на передачу 1 бита, то, очевидно, $R=1/T$, $h^2=(P_cT\log m)/N_0$, $F\approx \approx m/(T\log m)$ для ЧМ и $F\approx 1/(T\log m)$ для ФМ или ОФМ [6]. Из рис. 7.9 следует, что в системах с ЧМ при увеличении числа позиций m энергетическая эффективность β увеличивается, а частотная эффективность γ уменьшается. В системах с ФМ и ОФМ, наоборот, с увеличением m коэффициент β уменьшается, а γ увеличивается. Таким образом, условия обмена β на γ из-за изменения числа сигналов с ЧМ и ФМ различные.

При частотно-фазовой модуляции (ЧФМ) достигается некоторый компромисс в получении необходимых значений β и γ. Элементарным сигналом при ЧФМ является колебание с одиой из n частот, фаза которого принимает одно из m фиксированных значений. Общее число сигналов в ансамбле равно $M=nm$. На кривой для ЧФМ первая цифра указывает число частотных позиций, а вторая — число позиций фазы.

В табл. 7.2 приведены результаты расчета выигрыша g и информационной эффективности η для некоторых систем передачи непрерывных сообщений при

Таблица 7.2

Система модуляпии	$\alpha=\frac{F}{F_c}$	$g=\frac{\rho_{вых}}{\rho_{вх}}$	$g'=\frac{g}{\alpha}$	$\eta=\frac{R}{C}$
АМ	2	0,2	0,1	0,42
БМ	2	2	1	0,50
ОМ	1	1	1	1
ФМ	20	222	11,5	0,12
ЧМ	20	666	33,3	0,17
ФИМ-АМ	20	666	33,3	0,17
ИКМ-АМ	20	250	12,5	0,23
ИКМ-ЧМ	20	500	25	0,32
ИКМ-ФМ	20	1000	50	0,48
ИС	20	6310	315	1

заданном зиачении $\rho_{вых}=40$ дБ. В качестве меры скорости передачи для непрерывных сообщеиий примем эпсилон-производительность гауссовского источника, равную $F_c\log\rho_{вых}$. Для таких систем формулу (7.28) можно записать в следующем виде:

$$\eta=\frac{\log\rho_{вых}}{\alpha\log(\rho_{вых}/g+1)}. \qquad (7.32)$$

Выигрыш g и обобщенный выигрыш g' рассчитывались по соответствующим формулам гл. 6 и 7 для оптимального приема сигналов. При этом полагалось, что во всех системах передается одно и то же сообщение с наивысшей частотой F_c и пикфактором $\Pi=3$; параметры систем оптимальные. Для сравнения приведены также результаты для идеальной системы (ИС).

Анализ полученных данных показывает, что наибольшая информационная эффективность связи достигается в однополосной системе. Однако помехоустойчивость этой системы (величина выигрыша), так же как и систем БМ и АМ, сравнительно низкая. Порог помехоустойчивости в системе ОМ отсутствует, а в системе АМ выражен слабо (см. рис. 6.13).

Одноканальные системы ЧМ и ФИМ примерно равноценны. В этих системах, а также в цифровых системах с ИКМ высокая помехоустойчивость может быть достигнута за счет увеличения ширины спектра сигнала, т. е. за счет частотной избыточности. Во всех этих системах резко выражен порог помехоустойчивости.

На рис. 7.10 приведены кривые эффективности $\beta=f(\gamma)$ для аналоговых и цифровых систем передачи непрерывных сообщений. Скорость R определялась по формуле для эпсилон-производительности $H'_\varepsilon = F_c \log \rho_{вых}$, где $\rho_{вых}$ в аналоговых системах принимались равными 30 или 40 дБ.

Для аналоговых систем цифры на кривых указывают значения α, а цифры в скобках — значения $\rho_{вых}$. В пороговой области кривые для ЧМ изображены пунктиром. Работу в этой области можно реализовать при следящем приеме (см. § 6.10).

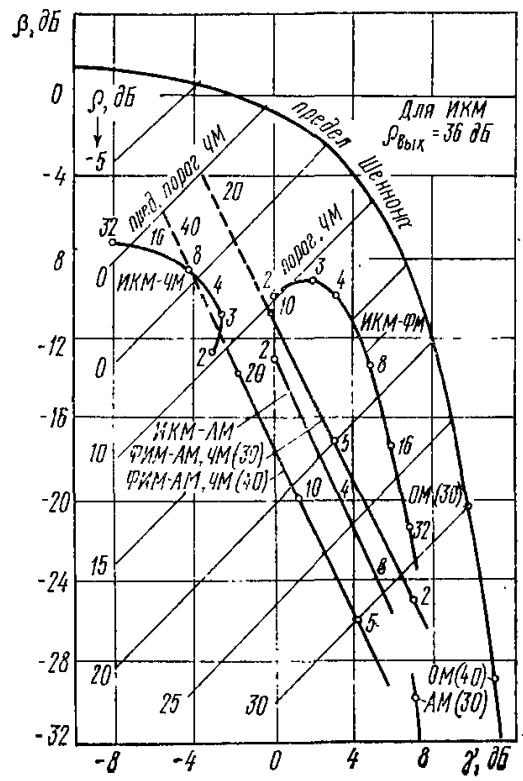


Рис. 7.10. Кривые энергетической и частотной эффективности некоторых систем передачи непрерывных сообщений

Для цифровых систем расчеты проводились для случая минимально необходимых полос канала F. Практически это соответствует каналу, в котором межсимвольные помехи полностью скорректированы. Здесь цифры на кривых указывают основание кода. Кривые рассчитаны для вероятности ошибки $p=10^{-5}$, что соответствует отношению сигнала к шуму $\rho_{вых}\approx 36$ дБ при равенстве мощностей шума квантования и шума ложных импульсов $\overline{\varepsilon^2_{кв}}=\overline{\varepsilon^2_{л}}$.

Из рис. 7.9 и 7.10 видно, что эффективность реальных систем существенно ниже предела Шеннона. Характер обмена между β и γ (ход кривых эффективности) зависит от вида модуляции и кода.

Приведенные номограммы на плоскости $\beta\gamma$ позволяют сравнительно быстро определить системы, удовлетворяющие заданным требованиям по энергетической и частотной эффективности.

В системах космической связи определяющим является необходимость обеспечить наилучшее использование мощности сигнала при заданной верности передачи. Из простых сигналов этому условию наиболее полно удовлетворяют дискретные системы с ФМ и ОФМ. Наиболее эффективна четырехфазная система, относящаяся к классу биортогональных. Эта система обеспечивает высокую энергетическую эффективность β при достаточно узкой полосе частот канала ($\gamma=2$).

Аналоговые системы ОМ, АМ и узкополосная ЧМ обеспечивают высокую частотную эффективность γ при низкой энергетической эффективности β. Применение этих систем целесообразно в каналах с хорошей энергетикой (при больших значениях $\rho_{вх}$) или в тех случаях, когда требуемое значение $\rho_{вых}$ мало. Цифровые системы обеспечивают высокую β-эффективность при сравнительно хорошей γ-эффективности. В каналах с ограниченной энергетикой (при малых значениях $\rho_{вх}$) преимущества цифровых систем особенно заметны. При высоком качестве передачи, когда требуемые значения $\rho_{вых}$ велики, широкополосная ЧМ и цифровые системы обеспечивают примерно одинаковую эффективность.

В ряде случаев при жестком ограничении мощности сигнала целесообразно произвести обмен γ на β (полосы пропускания канала на мощность сигнала). Это можно сделать при дискретной передаче путем применения сложных сигналов с основанием $m>2$ или корректирующих кодов, а в аналоговых системах путем увеличения α вплоть до величины, определяемой порогом помехоустойчивости. Переход к сложным сигналам эффективен при малых отношениях сигнала к шуму, в то время как корректирующие коды более эффективны при сравнительно больших отношениях сигнала к шуму. Поэтому практически могут оказаться целесообразными системы, в которых используются сложные сигналы, уменьшающие вероятность ошибок, которые затем

исправляются корректирующим кодом. Это фактически системы с двумя ступеиями кодироваиия. При этом могут примеияться различные корректирующие коды. Для космических систем уже в иастоящее время перспективнымн являются сверточные коды.

В системах проводной связи важнейшим показателем является частотная эффективность γ. Здесь определяющим является требование, согласно которому иеобходимо добиться наилучшего использования полосы пропускания канала при заданной верности передачи. Этому условию наиболее полно отвечает однополосная модуляция (ОМ).

Проблема эффективного использования полосы частот важна ие только для систем проводной связи, но и для радиосистем, в частности для спутниковых систем связи. Для повышения γ-эффективности можно, прежде всего, рекомендовать применение миогоуровневых и миогофазных сигналов, позволяющих улучшить использование полосы частот за счет некоторого снижения β-эффективности. Определенный выигрыш в ширине спектра дают также комбинированные виды модуляции, такие, как амплитудно-фазовая (АФМ), частотно-фазовая (ЧФМ) и т. п.

Можно рассчитывать на существенный выигрыш как в β-, так и γ-эффективности, если вместо двоичных применять коды с более высоким основанием. Это направление, по-видимому, является основным на путях приближения к пределу Шеннона. Однако следует помнить, что по мере приближения R к C (к пределу Шеннона) сложиость системы будет неизбежно возрастать и наступит такая ситуация, когда дальнейшее приближение к пределу стаиет нецелесообразным. Здесь важно определить тот практический предел, к которому следует стремиться и который оправдан техническими и экономическими соображениями.

В реальных условиях эффективность систем связи существенно снижается из-за потерь информации в канале. Основными причинами этих потерь являются межсимвольные и межканальные помехи, неточность формирования и сиихронизации сигналов, нестабильность частоты и т. п. Поэтому оптимизация параметров выбранного варианта системы должна производиться с учетом этих факторов.

Все приведенные результаты получены в предположении, что в принимаемом сигнале все параметры, кроме модулируемого, полностью известны. Это зиачит, что в канале имеется только аддитивная помеха. Для канала с неопределенной фазой, а тем более с замираниями, эффективность и выигрыш уменьшаются. При неопределеиной фазе приходится либо применять некогерентный прием (например, при АМ использовать детектор огибающей вместо синхронного детектора), либо оценивать фазу по ранее принятым отрезкам сигнала (адаптивный квазикогерентный прием).

Не следует думать, что рассмотренные в этом параграфе показатели эффективности полностью определяют целесообразность примеиения той или иной системы модуляции. Так, например, здесь совершенно не учитывались требования для систем с ретрансляторами. Возможность применеиия регеиераторов без накопления шумов квантования в цифровых системах (см. § 7.1) позволила существенно расширить спектр сигналов, передаваемых по кабелю. Поэтому системы с ИКМ в ряде случаев оказываются выгоднее, чем ОМ, несмотря на их меньшую частотную эффективность.

Относительность понятия эффективности ярко проявилась на примере радиовещания. Казалось бы, что АМ для этого совершенно непригодна, так как для иее β значительно меньше, чем для любой другой модуляпии, а γ меньше, чем для ОМ. Тем не менее, в радиовещании на длинных, средних и коротких волнах используется только АМ. Это объясняется очень просто. В отличие от систем радиосвязи, где на каждый передатчик приходится примерио один приемник, при радиовещании передатчик обслуживает десятки тысяч приемников. Поэтому для обеспечения удовлетворительной верности экономическн выгоднее в несколько раз увеличить мощность передатчика, нежели значительиио усложиить приемник, применив в нем вместо простого детектора огибающей, состоящего из диода, пары резисторов и кондеисаторов, иапример сиихроииый детектор с автоподстройкой, необходимый при ОМ или БМ. Впрочем, когда требования к качеству вещания возрастают, АМ становится иевыгодиой, поскольку стоимость передатчика быстро возрастает с его мощностью.

Поэтому высококачественное радиовещание использует частотную модуляцию и ведется на УКВ, где можно занять достаточную полосу частот, хотя при этом и сокращается радиус действия вещательного передатчика.

ГЛАВА ВОСЬМАЯ

ТЕОРИЯ МНОГОКАНАЛЬНОЙ ПЕРЕДАЧИ СООБЩЕНИЙ

8.1. ОСНОВЫ ТЕОРИИ РАЗДЕЛЕНИЯ СИГНАЛОВ

Непрерывное развитие народного хозяйства и культуры приводит к непрерывному росту потоков передаваемой информации. А следовательно, необходимо постоянно увеличивать число каналов передачи и их пропускную способность.

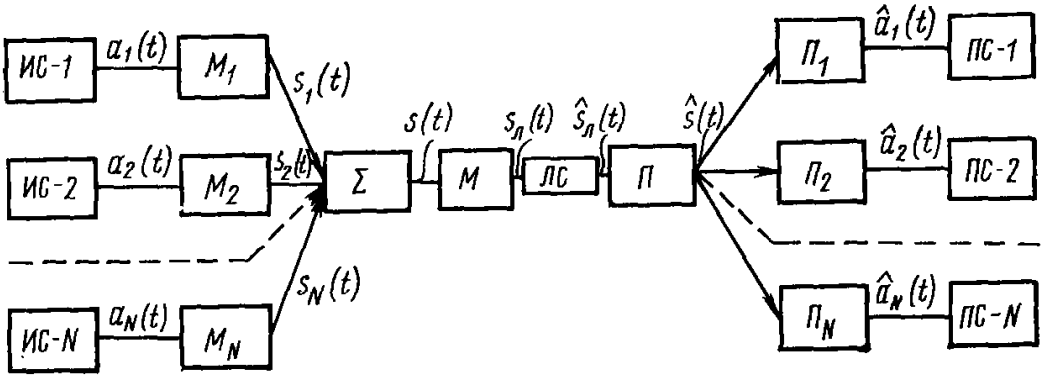


Рис. 8.1. Структурная схема линейной многоканальной системы связи

Практика построения современных сетей передачи информации показывает, что наиболее дорогостоящими звеньями каналов являются линии связи (кабельные, волноводные и световодные линии, линии радиорелейной и спутниковой связи и др.). Поскольку экономически нецелесообразно использовать дорогостоящую линию связи для передачи информации единственной пары абонентов (источника и получателя), то возникает задача построения многоканальных систем передачи, в которых общая линия «уплотняется» большим числом индивидуальных каналов. Этим обеспечивается повышение эффективности использования пропускной способности линии. Разумеется, многоканальная передача возможна в тех случаях, когда пропускная способность линии C не меньше суммарной производительности «уплотняемых» независимых источников информации: $C \geqslant \sum_{k=1}^{N} H'_k$, где H'_k — производительность k-го источника, а N — число источников (каналов), называемое также кратностью уплотнения.

С целью унификации многоканальных систем связи за основной или стандартный канал принимают канал тональной частоты

(канал ТЧ), обеспечивающий передачу сообщений с эффективно передаваемой полосой частот 300—3400 Гц, соответствующей основному спектру телефонного сигнала.

Многоканальные системы образуются путем объединения каналов ТЧ в группы, обычно кратные 12 каналам. В свою очередь, часто используется «вторичное уплотнение» каналов ТЧ телеграфными каналами и каналами передачи цифровой информации (каналами передачи данных).

На рис. 8.1 приведена структурная схема наиболее распространенных линейных систем многоканальной связи. Реализации сообщений каждого источника $a_1(t)$, $a_2(t)$, ..., $a_N(t)$ с помощью индивидуальных передатчиков (модуляторов) M_1, M_2, ..., M_N преобразуются в соответствующие канальные сигналы $s_1(t)$, $s_2(t)$, ..., $s_N(t)$. Совокупность канальных сигналов на выходе суммирующего устройства Σ образует *групповой сигнал* $s(t)$. Наконец, в групповом передатчике M сигнал $s(t)$ преобразуется в линейный сигнал $s_л(t)$, который и поступает в линию связи ЛС. Простоты ради, считаем пока, что линия пропускает сигнал без искажений и не вносит шумов. Тогда на приемном конце линии связи линейный сигнал $\hat{s}_л(t)$ с помощью группового приемника $П$ может быть вновь преобразован в групповой сигнал $\hat{s}(t)$. Канальными или индивидуальными приемниками $П_1$, $П_2$, ..., $П_N$ из группового сигнала $s(t)$ выделятся соответствующие канальные сигналы $s_1(t)$, $s_2(t)$, ..., $s_N(t)$ и затем преобразуются в предназначенные получателям сообщения $\hat{a}_1(t)$, $\hat{a}_2(t)$, ..., $\hat{a}_N(t)$.

Канальные передатчики вместе с суммирующим устройством образуют аппаратуру уплотнения. Групповой передатчик M, линия связи ЛС и групповой приемник $П$ составляют групповой канал связи, который вместе с аппаратурой уплотнения и индивидуальными приемниками объединяется в систему многоканальной связи.

Индивидуальные приемники системы многоканальной связи $П_k$ наряду с выполнением обычной операции преобразования сигналов $s_k(t)$ в соответствующие сообщения $a_k(t)$ должны обеспечить выделение сигналов $s_k(t)$ из группового сигнала $s(t)$. Иначе говоря, в составе технических устройств на передающей стороне многоканальной системы должна быть предусмотрена аппаратура уплотнения, а на приемной стороне — аппаратура разделения.

В общем случае групповой сигнал может формироваться не только простейшим суммированием канальных сигналов, но также и определенной логической обработкой, в результате которой каждый элемент группового сигнала несет информацию о сообщениях источников. Это так называемые системы с комбинационным уплотнением, которые будут рассмотрены в § 8.5.

Перейдем теперь к вопросу об общих свойствах сигналов, пригодных для одновременной и независимой передачи в системах многоканальной связи. Чтобы разделяющие устройства были в состоянии различать сигналы отдельных каналов, долж-

ны существовать определенные признаки, присущие только данному сигналу. Такими признаками в общем случае могут быть параметры переносчика, например амплитуда, частота или фаза в случае непрерывной модуляции синусоидального переносчика. При дискретных видах модуляции различающим признаком может служить и форма сигналов. В соответствии с используемым для разделения признаком различаются и способы разделения сигналов: частотный, временной, фазовый и др.

Пусть, например, необходимо организовать одновременную работу N индивидуальных каналов по общему групповому каналу. Будем считать, что групповой канал пригоден для передачи сигналов любого k-го канала $s_k(t)$. Предположим, что сигнал k-го канала можно представить в виде

$$s_k(t) = C_k \psi_k(t), \tag{8.1}$$

где ψ_k — функция переносчика; C_k — некоторый коэффициент, отображающий передаваемое сообщение (при непрерывных сообщениях C_k означает мгновенное значение функции сообщения; при дискретной передаче C_k — некоторое число, соответствующее передаваемому символу).

Для суммы всех канальных сигналов (группового сигнала) можем записать:

$$s(t) = \sum_{k=1}^{N} s_k(t) = \sum_{k=1}^{N} C_k \psi_k(t). \tag{8.2}$$

Групповой сигнал затем преобразуется в линейный сигнал $s_л(t)$, который и передается в линию связи. На приемном конце линейный сигнал $s_л(t)$ вновь преобразуется в групповой сигнал, т. е. преобразуется к виду $s(t)$, удобному для выполнения операции разделения сигналов.

Для разделения N канальных сигналов на приемной стороне группового канала необходимо иметь N разделяющих устройств, причем каждое k-е разделяющее устройство должно выполнять операцию выделения k-го сигнала. Действие приемного устройства, в результате которого происходит выделение сигналов определенного k-го канала, будем для краткости условно обозначать оператором разделения π_k. Приемное устройство, описываемое оператором разделения π_k, только тогда полностью выделит сигнал $s_k(t)$, когда оно совершенно не будет реагировать на сигналы других каналов. Другими словами, идеальное k-е приемное устройство должно реагировать («откликаться») только на сигнал $s_k(t)$ и не должно откликаться на остальные сигналы. Наложим на оператор π_k дополнительное условие, потребовав, чтобы он был линейным. Это значит, что он должен удовлетворять принципу суперпозиции:

$$\pi_k\{s_1(t) + s_2(t)\} = \pi_k\{s_1(t)\} + \pi_k\{s_2(t)\}. \tag{8.3}$$

Теперь уже легко сформулировать операцию разделения сиг-

налов в математическом виде. Обозначим через $\hat{s}_k(t)$ отклик, т. е. результат воздействия оператора π_k приемного устройства k-го канала на групповой сигнал $s(t)$, т. е. $\pi_k\{s(t)\} = \hat{s}_k(t)$.

На входе каждого k-го приемного устройства многоканальной системы действует сумма сигналов всех N каналов. Для того чтобы приемное устройство $П_k$ было «чувствительным» только по отношению к сигналам $s_k(t)$, необходимо, чтобы его отклики на все другие сигналы были равны нулю:

$$u_k(t) = \pi_k\left\{\sum_{i=1}^{N} s_i(t)\right\} = \sum_{i=1}^{N} \pi_k\{s_i(t)\} = s_k(t).$$

Здесь второе равенство вытекает из линейности оператора π_k.

Очевидно, что для этого достаточно выполнить для всех i и k условие

$$\pi_k\{s_i(t)\} = \begin{cases} s_k(t) & i = k; \\ 0 & i \neq k. \end{cases} \tag{8.4}$$

Подставляя сюда (8.1), получим

$$\pi_k\{C_i \psi_i(t)\} = \begin{cases} C_k \psi_k & i = k; \\ 0 & i \neq k \end{cases}$$

и, следовательно, $\hat{s}_k(t) = C_k \psi(t)$.

Полученные результаты могут быть обобщены также на случай, когда отклик разделяющего устройства на сигнал $s_k(t)$ будет иметь иную форму; важно лишь, чтобы величина отклика была однозначно связана с передаваемым сигналом. В частном случае, откликом на сигиал $s_k(t)$ может быть просто некоторое число d_k, однозначно связанное с коэффициентом C_k, например пропорциональное C_k:

$$\hat{s}_k = \pi_k\{s(t)\} = \pi_k\left\{\sum_{i=1}^{N} C_i \psi_i(t)\right\} = \sum_{i=1}^{N} \pi_k\{C_i \psi_i(t)\} = d_k \tag{8.5}$$

или

$$\pi_k\{C_i \psi_i(t)\} = \begin{cases} d_k & i = k; \\ 0 & i \neq k. \end{cases}$$

Физический смысл полученных выражений (8.4) и (8.5) сводится к тому, что приемное устройство $П_k$ выделяет только «свои» сигналы $s_k(t)$ и не реагирует на сигналы всех других каналов, т. е. приемник $П_k$ обладает избирательными свойствами по отношению к сигналам $s_k(t)$. Поскольку действие приемников $П_k$ в формулах (8.4)—(8.6) описывается линейным оператором π_k, то и соответствующие устройства разделения называются линейными. Напомним, что здесь рассматривался случай идеального разделения, когда отклик k-го разделяющего устройства на сигналы всех других каналов равен нулю. В реальных условиях при разделении сигналов возникают *переходные помехи.*

Как известно, воздействия любого линейного оператора π_k на функцию $s(t)$ можно представить в виде скалярного произведения (интегрального преобразования);

$$\pi_k\{s(t)\} = \int_{-\infty}^{\infty} s(\tau)\,\eta_k(t,\tau)\,d\tau, \tag{8.6}$$

где $\eta_k(t, \tau)$ — некоторая *весовая функция*, соответствующая оператору π_k.

Впервые определение избирательных свойств приемника в соответствии с формулами (8.4) — (8.6) было предложено в 1935 г. Д. В. Агеевым, который доказал, что необходимым и достаточным условием разделимости сигналов (8.2) линейными устройствами является условие линейной независимости. Оно заключается в том, что тождество

$$C_1\psi_1(t) + C_2\psi_2(t) + \ldots + C_k\psi_k(t) + \ldots + C_N\psi_N(t) \equiv 0 \tag{8.7}$$

может выполняться в том единственном случае, когда все коэффициенты C_k одновременно равны нулю (см. § 2.5).

Действительно, в соответствии с определением избирательных свойств (8.4) и (8.5) подлежащие разделению сигналы $s_k(t)$ и линейные приемные устройства Π_k в общем виде должны удовлетворять условию линейного разделения:

$$\pi_k\{\psi_i(t)\} = \gamma_{ik}, \quad i, k = 1, 2, \ldots, N, \tag{8.8}$$

где γ_{ik} — отклик разделяющего устройства на канальный сигнал $s_i(t)$, причем $\gamma_{ik} = 0$ при $i \neq k$ и $\gamma_{kk} \neq 0$. Если теперь подействовать оператором π_k на обе части тождества (8.7), то, принимая во внимание (8.8) и принцип суперпозиции, получим

$$\pi_k\left\{\sum_{i=1}^{N} C_i\psi_i(t)\right\} = \sum_{i=1}^{N} C_i\,\pi_k\{\psi_i(t)\} = C_k\gamma_{kk} = 0. \tag{8.9}$$

Но функции γ_{kk} не равны тождественно нулю, следовательно, равны нулю все C_k, $k=1, 2, \ldots, N$. Иначе говоря, условие линейного разделения (8.4) будет выполняться лишь тогда, когда канальные сигналы линейно независимы. Частным случаем линейно независимых сигналов, как было показано в § 2.5, являются ортогональные сигналы.

Напомним, что в n-мерном евклидовом пространстве, в частности в пространстве сигналов, представимых ограниченным рядом Котельникова, для которых $n=2FT$ (2.117), можно выбрать не более n линейно независимых сигналов.

На геометрическом языке векторы $\psi_1, \psi_2, \ldots, \psi_N$ линейно независимы тогда и только тогда, когда ни один из них не может быть образован линейной комбинацией других.

Условию разделения (8.8) можно придать отчетливую геометрическую трактовку, если воспользоваться определениями линейного подпространства и оператора проектирования (см. § 2.5). Пусть $\{\psi_1, \psi_2, \ldots, \psi_n\}$ — базис в n-мерном пространстве Y_n. Поставим в соответствие вектору $y = \sum_{k=1}^{n} \lambda_k\psi_k$ вектор

$x = \mathcal{P}_{nm}\{y\} = \sum_{k=1}^{m} \lambda_k \psi_k$, где $m < n$. Оператор $\mathcal{P}_{nm}$ называется оператором проектирования пространства Y_n на подпространство X_m с базисом $\{\psi_1, \psi_2, ..., \psi_m\}$.

Выберем ансамбли канальных сигналов $s_1, s_2, ..., s_N$ так, чтобы они составляли непересекающиеся подпространства пространства группового сигнала $\mathcal{S}$.

Задача выделения сигналов $s_k(t)$ k-го канала, сводится, по существу, к операции проектирования пространства $\mathcal{S}$ на подпространство $\mathcal{S}_k$ с помощью оператора проектирования $\mathcal{P}$, роль которого в нашем случае выполняет оператор π_k:

$$\pi_k\{s\} = \pi_k\left\{\sum_{i=1}^{N} s_i\right\}. \qquad (8.10)$$

Таким образом, для разделения сигналов $s_1, s_2, ..., s_N$ необходимы такие устройства, чтобы описывающие их операторы являлись операторами проектирования, а подпространства $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_N$ должны быть непересекающимися. Разделенные таким образом сигналы затем преобразуются демодуляторами в сообщения $a_1, a_2, ..., a_N$.

Заметим, что теория линейного разделения построена без учета аддитивных помех. В этих условиях безразлично, выбирать ли систему переносчиков $\{\psi\}$ ортогональных или просто линейно независимых. В любом случае их максимальное число равно размерности пространства $2FT$. Однако если учесть влияние помех, то преимущество ортогональных сигналов неоспоримо, так как только в этом случае энергия сигнала используется полностью. Поэтому, как правило, переносчики выбираются ортогональными. Частным случаем ортогональных сигналов являются функции с неперекрывающимися спектрами, а также функции, не перекрывающиеся во времени.

8.2. ЧАСТОТНОЕ, ВРЕМЕННОЕ И ФАЗОВОЕ РАЗДЕЛЕНИЯ СИГНАЛОВ

Частотное разделение сигналов. Функциональная схема простейшей системы многоканальной связи с разделением каналов по частоте представлена на рис. 8.2.

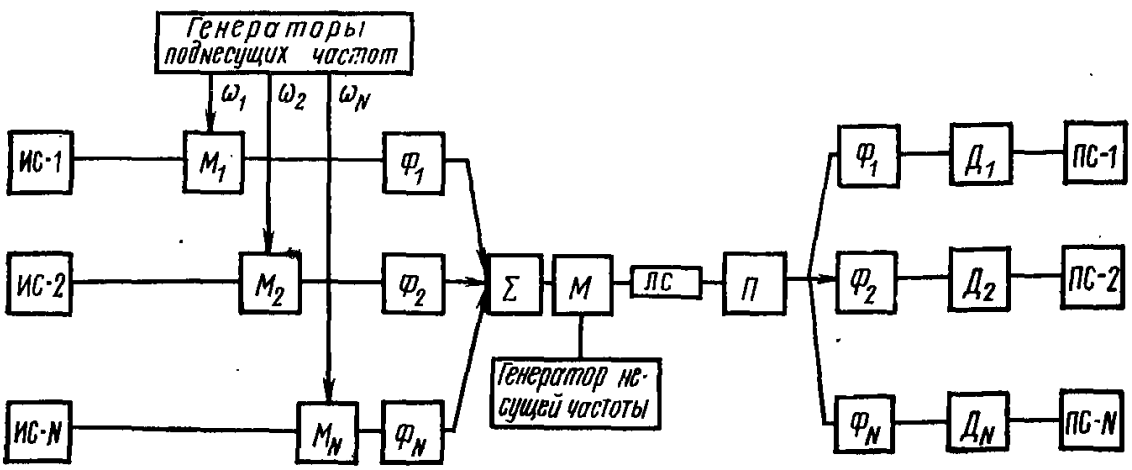


Рис. 8.2. Схема многоканальной системы с частотным разделением

Проследим основные этапы образования сигналов, а также изменение этих сигналов в процессе передачи. Сначала в соответствии с передаваемыми сообщениями первичные (индивидуальные) сигналы, имеющие энергетические спектры $G_1(\omega)$, $G_2(\omega)$, ..., $G_N(\omega)$, модулируют поднесущие частоты ω_k каждого канала. Эта операция выполняется модуляторами M_1, M_2, ..., M_n канальных передатчиков. Полученные на выходе частотных фильтров Φ_1, Φ_2, ..., Φ_N спектры $g_k(\omega)$ канальных сигналов занимают соответственно полосы частот $\Delta\omega_1$, $\Delta\omega_2$, ..., $\Delta\omega_N$ (рис. 8.3), которые в общем случае могут отличаться по ширине от

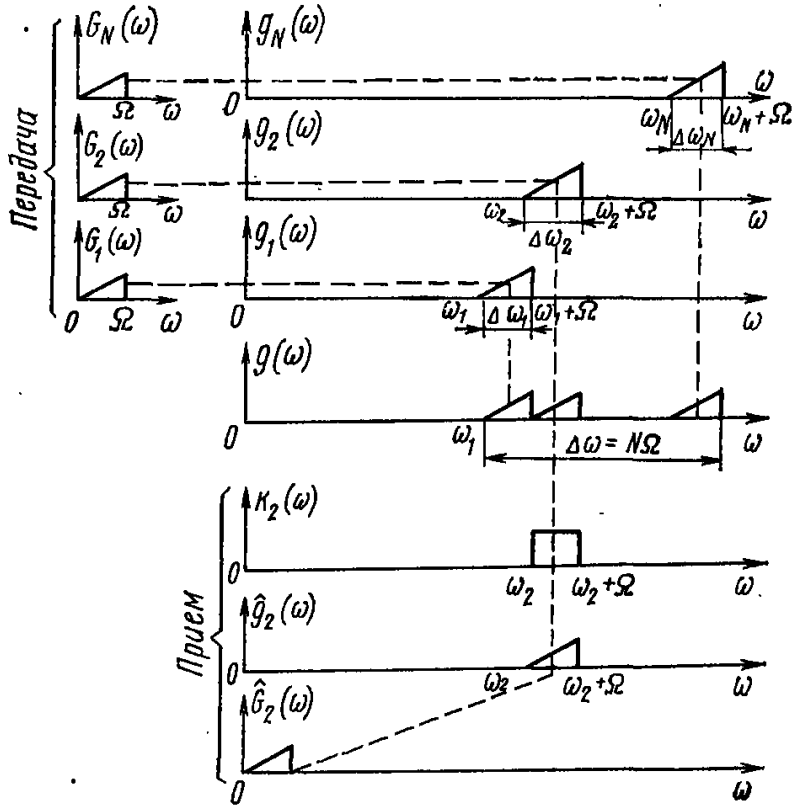


Рис. 8.3. Преобразование спектров в системе с частотным разделением сигналов

спектров сообщений Ω_1, Ω_2, ..., Ω_N. При широкополосных видах модуляции, например ЧМ, ширина спектра $\Delta\omega_k \approx 2(\beta+1)\Omega_k$, при ОМ $\Delta\omega_k = \Omega_k$, т. е. в общем случае $\Delta\omega_k \geqslant \Omega_k$. Простоты ради будем считать, что используется ОМ (как это принято в системах многоканальной связи с частотным разделением), т. е.

$$\Delta\omega_k = \Omega \quad \text{и} \quad \Delta\omega = N\Omega. \tag{8.11}$$

Будем полагать, что спектры индивидуальных сигналов финитны. Тогда можно подобрать поднесущие частоты ω_k так, что полосы $\Delta\omega_1$, ..., $\Delta\omega_k$ попарно не перекрываются. При этом условии сигналы $s_k(t)$ $(k=1, \ldots, N)$ взаимно ортогональны. Затем спектры $g_1(\omega)$, $g_2(\omega)$, ..., $g_N(\omega)$ суммируются (СУ) и их совокупность $g(\omega)$ поступает на групповой модулятор M. Здесь спектр $g(\omega)$ с помощью колебания несущей частоты ω_0 переносится в область

частот, отведенную для передачи данной группы каналов, т. е. групповой сигнал $s(t)$ преобразуется в линейный сигнал $s_л(t)$. При этом может использоваться любой вид модуляции.

На приемном конце линейный сигнал поступает на групповой демодулятор (приемник $П$), который преобразует спектр линейного сигнала в спектр группового сигнала $g(\omega)$. Спектр группового сигнала затем с помощью частотных фильтров $Ф_1$, $Ф_2$, ..., $Ф_N$ вновь разделяется на отдельные полосы $\Delta\omega_k$, соответствующие отдельным каналам. Наконец, канальные демодуляторы $Д$ преобразуют спектры сигналов $\hat{g}_k(\omega)$ в спектры сообщений $G_k(\omega)$, предназначенные получателям.

Из приведенных пояснений легко понять смысл частотного способа разделения каналов. Поскольку всякая реальная линия связи обладает ограниченной полосой пропускания, то при многоканальной передаче каждому отдельному каналу отводится определенная часть общей полосы пропускания.

На приемной стороне одновременно действуют сигналы всех каналов, различающиеся положением их частотных спектров на шкале частот. Чтобы без взаимных помех разделить такие сигналы, приемные устройства должны содержать частотные фильтры. Каждый из фильтров $Ф_k$ должен пропустить без ослабления лишь те частоты $\omega \in \Delta\omega_k$, которые принадлежат сигналу данного канала; частоты сигналов всех других каналов $\omega \notin \Delta\omega_k$ фильтр должен подавить. Такой способ разделения сигналов называется частотным разделением.

Математически частотное разделение сигналов идеальными полосовыми фильтрами можно представить следующим образом:

$$\hat{s}_k(t) = \int_{-\infty}^{\infty} s(\tau)\, g_k(t-\tau)\, d\tau, \qquad (8.12)$$

где $g_k(t)$ — импульсная реакция идеального полосового фильтра, пропускающего без искажений полосу частот $\Delta\omega_k$. Выражение (8.12) совпадает с (8.6) при весовой функции $\eta_k(t, \tau) = g_k(t-\tau)$. В спектральной области преобразование (8.12) соответствует умножению спектра группового сигнала на П-образную передаточную функцию (см. рис. 8.3).

Итак, с точки зрения возможности полного разделения сигналов различных каналов надо иметь такие фильтры $Ф_k$, полоса пропускания которых полностью соответствует ширине спектра сигнала $\Delta\omega_k$; на гармонические составляющие за пределами полосы $\Delta\omega_k$ фильтр $Ф_k$ реагировать не должен. При этом имеется в виду, что энергия сигналов s_k полностью сосредоточена в пределах ограниченной полосы $\Delta\omega_k$, отведенной k-му каналу. Если бы оба эти условия удовлетворялись, то посредством частотных фильтров можно было бы разделить сигналы различных каналов без взаимных помех. Однако ни одно из этих условий принципиально невыполнимо. Результатом являются взаимные помехи между каналами. Они возникают как за счет неполного сосредоточения энергии сигнала k-го канала в пределах заданной

полосы частот $\Delta\omega_k$, так и за счет неидеальности реальных полосовых фильтров. В реальных условиях приходится учитывать также взаимные помехи нелинейного происхождения, например за счет нелинейности характеристик группового канала.

Для снижения переходных помех до допустимого уровня приходится вводить защитные частотные интервалы $\Delta\omega_{защ}$ (рис. 8.4). Так, например, в современных системах многоканальной телефонной связи каждому телефонному каналу выделяется полоса частот 4 кГц, хотя частотный спектр передаваемых звуковых сигналов ограничивается полосой от 300 до 3400 Гц, т. е. ши-

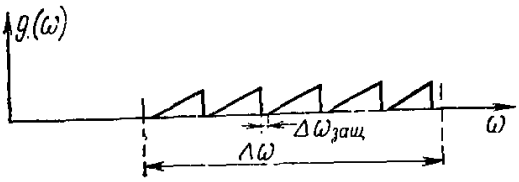


Рис. 8.4. Спектр группового сигнала с защитными промежутками

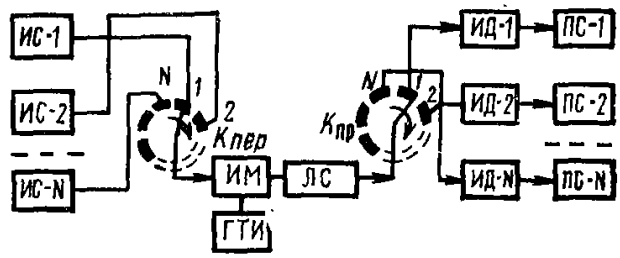


Рис. 8.5. Схема многоканальной системы с временным разделением

рина спектра составляет 3,1 кГц. Между полосами частот соседних каналов предусмотрены интервалы шириной по 0,9 кГц, предназначенные для снижения уровня взаимных помех при расфильтровке сигналов. Это означает, что в многоканальных системах связи с частотным разделением сигналов эффективно используется лишь около 80% полосы пропускания линии связи. Кроме того, необходимо обеспечить очень высокую степень линейности всего тракта группового сигнала.

Временно́е разделение сигналов. Принцип временно́го разделения сигналов весьма прост и издавна применяется в телеграфии. Он состоит в том, что с помощью коммутатора $K_{пер}$ групповой тракт предоставляется поочередно для передачи сигналов каждого канала многоканальной системы[1]. При передаче непрерывных сообщений для временно́го уплотнения используется дискретизация по времени (импульсная модуляция). Сначала передается сигнал (импульс) 1-го канала, затем следующего канала и т. д., до последнего канала за номером N, после чего опять включается 1-й канал и процесс периодически повторяется (рис. 8.5).

На приемном конце устанавливается аналогичный коммутатор $K_{пр}$, который подключает групповой тракт поочередно к приемникам разных каналов. Приемник каждого k-го канала должен быть подключен только на время передачи k-го сигнала и выключен все остальное время, пока передаются сигналы в других каналах. Это означает, что для нормальной работы многоканальной системы с временны́м разделением необходима синхронная и синфазная работа коммутаторов на приемной и передающей сторонах. Часто с этой целью один из каналов занимается под передачу специаль-

[1] В современной аппаратуре механические коммутаторы практически не используются. Вместо них применяют электронные коммутаторы, выполненные, напрнмер, на регистрах сдвига.

ных импульсов синхронизации, предназначенных для согласованной во времени работы $K_{пер}$ и $K_{пр}$.

На рис. 8.6 представлены временны́е диаграммы двухканальной системы с АИМ. Переносчиком сообщений здесь являются последовательности импульсов (с периодом $T_0 = 1/2F_{макс}$), поступающих на импульсный модулятор (ИМ) от генератора тактовых импульсов (ГТИ). Групповой сигнал (рис. 8.6*а*) поступает на ком-

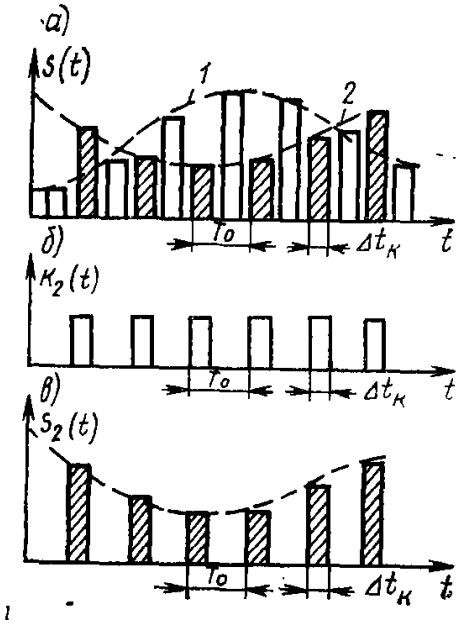


Рис. 8.6. Временно́е разделение двух сигналов с АИМ

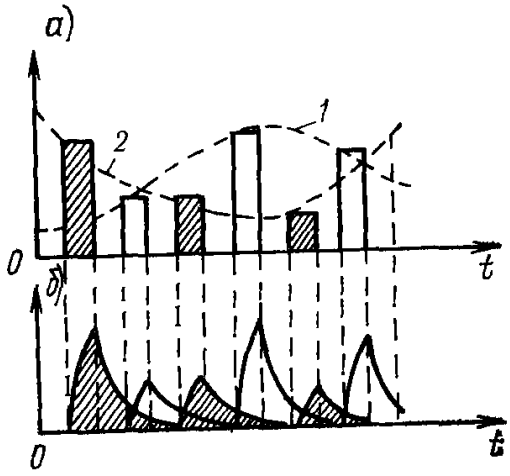


Рис. 8.7. Возникновение переходных помех при временно́м разделении

мутатор $K_{пр}$. Последний выполняет роль «временны́х» параметрических фильтров или ключей, передаточная функция которых K_k (рис. 8.6*б*) изменяется синхронно (с периодом T_0) и синфазно с изменениями передаточной функции $K_{пер}$:

$$K_k(t) = \begin{cases} 1 & t \in \Delta t_k; \\ 0 & t \notin \Delta t_k. \end{cases} \tag{8.13}$$

Это означает, что к тракту передачи в пределах каждого временно́го интервала Δt_k подключен только k-й импульсный детектор ИД-k. Полученные в результате детектирования $s_k(t)$ сообщения поступают к получателю сообщений ПС-k.

Оператор π_k, описывающий работу ключевого фильтра, вырезает из сигнала $s(t)$ интервалы Δt_k, следующие с периодом T_0, и отбрасывает остальную часть сигнала. Легко убедиться, что его можно представить в форме (8.6), если

$$\eta_k(t, \tau) = \begin{cases} \delta(t-\tau) & \text{при } t \in \Delta t_k; \\ 0 & \text{при } t \notin \Delta t_k. \end{cases} \tag{8.14}$$

Здесь, как и ранее, Δt_k обозначает интервал, в течение которого передаются сигналы k-го источника.

При временно́м разделении взаимные помехи в основном обусловлены двумя причинами. Первая из них состоит в том, что линейные искажения, возникающие за счет ограниченности полосы частот и неидеальности амплитудно-частотной и фазо-частотной характеристик всякой физически осуществимой системы связи, нарушают импульсный характер сигналов. Действительно, если при передаче модулированных импульсов конечной длительности ограничить спектр, то импульсы «расплывутся» и вместо импульсов конечной длительности мы получим процессы, бесконечно протяженные во времени. При временно́м разделении сигналов это приведет к тому, что импульсы одного канала будут накладываться на импульсы других каналов (рис. 8.7). Иначе говоря, между каналами возникают взаимные переходные помехи или, как их иногда называют, межсимвольная интерференция. Кроме того, взаимные помехи могут возникать за счет несовершенства синхронизации тактовых импульсов на передающей и приемной сторонах.

Для снижения уровня взаимных помех приходится вводить «защитные» временны́е интервалы, что соответствует некоторому расширению спектра сигналов. Так, в многоканальных системах коммерческой телефонии полоса эффективно передаваемых частот устанавливается равной $F=3100$ Гц; в соответствии с теоремой Котельникова минимальное значение $f_0=2F=6200$ Гц. Однако в реальных системах частоту следования импульсов выбирают с некоторым запасом и принимают равной $f_0=8$ кГц. Для передачи таких импульсов в одноканальном режиме потребуется полоса частот не менее 4 кГц. При временно́м разделении каналов сигнал каждого канала занимает одинаковую полосу частот, определяемую в идеальных условиях, согласно теореме Котельникова, из соотношения (без учета канала синхронизации)

$$\Delta t_k = \frac{T_0}{N} = \frac{1}{N2F} = \frac{1}{2F_{\text{общ}}}, \tag{8.15}$$

где $F_{\text{общ}}=NF$, что совпадает с общей полосой частот системы при частотном разделении. Хотя теоретически временно́е и частотное разделения позволяют достигнуть одинаковой эффективности использования частотного спектра, тем не менее пока что системы временно́го разделения уступают системам частотного разделения по этому показателю. Поэтому при большом числе каналов N и ограниченности используемой полосы частот предпочитают применять частотное разделение.

Вместе с тем, системы с временны́м разделением имеют неоспоримое преимущество, связанное с тем, что благодаря разновременности передачи сигналов разных каналов отсутствуют переходные помехи нелинейного происхождения. Кроме того, аппаратура временно́го уплотнения значительно проще, чем при частотном уплотнении, где для каждого индивидуального канала требуются соответствующие полосовые фильтры, которые трудно реализовать средствами микроэлектроники. Немаловажным преимуществом систем временно́го уплотнения является значительно

меньший пик-фактор. Временно́е уплотнение широко используется при передаче непрерывных сообщений с аналоговой импульсной модуляцией, а также с ИКМ.

Заметим также, что суммарная мощность $P_{общ}$ принимаемого сигнала $s(t)$, необходимая для обеспечения заданной верности в присутствии флуктуационных помех, как при частотном, так и при временно́м разделениях (а также при других, рассматриваемых ниже системах с линейным разделением) в идеальном случае в N раз больше, чем мощность P при одноканальной передаче с тем же видом модуляции $P_{общ}=NP$. Это легко понять, поскольку при сложении независимых сигналов их мощности складываются. В действительности из-за переходных помех верность приема в многоканальной системе при выполнении этого условия несколько ниже, чем в одноканальной системе. Увеличивая мощность сигнала в многоканальной системе, нельзя снизить воздействие переходных помех, поскольку при этом и мощность последних также возрастает, а в случае помех нелинейного происхождения растет даже быстрее, чем мощность сигнала.

Фазовое разделение сигналов. Рассмотрим теперь множество синусоидальных сигналов

$$\left.\begin{aligned} s_1(t) &= A_1 \sin(\omega_0 t + \varphi_1);\\ s_2(t) &= A_2 \sin(\omega_0 t + \varphi_2);\\ &\cdots\cdots\cdots\cdots\\ s_k(t) &= A_k \sin(\omega_0 t + \varphi_k);\\ &\cdots\cdots\cdots\cdots\\ s_N(t) &= A_N \sin(\omega_0 t + \varphi_N). \end{aligned}\right\} \qquad (8.16)$$

Здесь подлежащая передаче информация содержится в изменениях амплитуды A_k (амплитудная модуляция), несущая частота сигналов ω_0 — одна и та же, а сигналы различаются начальными фазами φ_k.

Среди множества N сигналов (8.16) лишь любые два сигнала являются линейно независимыми; любые $n>2$ сигналов оказываются линейно зависимыми. Это означает, что на одной несущей частоте ω_0 при произвольных значениях амплитуд A_i и A_k и фаз φ_i и φ_k можно обеспечить лишь двухканальную передачу.

На практике преимущественно используется значение $\Delta\varphi=\frac{\pi}{2}$:

$$s_1(t) = A_1 \sin\omega_0 t;$$

$$s_2(t) = A_2 \sin\left(\omega_0 t + \frac{\pi}{2}\right) = A_2 \cos\omega_0 t. \qquad (8.17)$$

При этом сигналы $s_1(t)$ и $s_2(t)$ ортогональны, что облегчает техническую реализацию системы и улучшает ее энергетические показатели.

8.3. РАЗДЕЛЕНИЕ СИГНАЛОВ ПО ФОРМЕ

Для разделения сигналов могут использоваться не только такие очевидные признаки, как частота, время и фаза. Наиболее общим признаком является форма сигналов. Различающиеся по форме сигналы могут передаваться одновременно и иметь перекрывающиеся частотные спектры, и тем не менее такие сигналы можно разделить, если выполняется условие их линейной независимости или условие ортогональности. Пусть в качестве переносчиков выбраны импульсы, последовательность которых образует, например, степенной ряд:

$$\psi_1 = 1, \quad \psi_2 = t, \quad \psi_3 = t^2, \quad \ldots \qquad (0 \leqslant t \leqslant T). \tag{8.18}$$

В предположении, что информация содержится в коэффициентах c_1; c_2; ..., для группового сигнала запишем

$$s(t) = c_1 1 + c_2 t + \ldots + c_N t^N. \tag{8.19}$$

Члены ряда (8.18) линейно независимы, и, следовательно, ни один из канальных сигналов $c_k t^k$ не может быть образован линейной суммой всех других сигналов. Это легко понять, обратив внимание на то, что многочлен от t вида (8.19) может быть тождественно равен нулю только в том случае, когда все его коэффициенты равны нулю.

Для разделения таких сигналов можно применить общий метод в соответствии с условием линейной независимости переносчиков (8.6). Так, при двухканальной передаче, имеем для интервала $(0 \leqslant t \leqslant T)$

$$s = s_1 + s_2 = c_1 1 + c_2 t;$$

если далее весовые функции выбрать удовлетворяющими условиям разделения:

$$\left.\begin{aligned} \eta_1(t) &= a_{11}\psi_1 + a_{12}\psi_2 = \frac{4}{T}\psi_1 - \frac{6}{T}\psi_2 = \frac{4}{T} - 6\frac{t}{T}, \\ \eta_2(t) &= a_{21}\psi_1 + a_{22}\psi_2 = -\frac{6}{T}\psi_1 + \frac{12}{T}\psi_2 = -\frac{6}{T} + 12\frac{t}{T} \quad (0 \leqslant t \leqslant T), \end{aligned}\right\} \tag{8.20}$$

то в результате операций проектирования будем иметь

$$\left.\begin{aligned} \pi_1[s] &= \int_0^T s(t)\,\eta_1(t)\,dt = 4c_1(\psi_1\psi_1) - 6c_1(\psi_1\psi_2) + 4c_2(\psi_2\psi_1) - \\ &\qquad - 6c_2(\psi_2\psi_2) = c_1; \\ \pi_2[s] &= \int_0^T s(t)\,\eta_2(t)\,dt = -6c_1(\psi_1\psi_2) + 12c_1(\psi_1\psi_1) - \\ &\qquad - 6c_2(\psi_1\psi_2) + 12c_2(\psi_2\psi_2) = c_2 \end{aligned}\right\} \tag{8.21}$$

[здесь $(\psi_1\psi_1) = 1$, $(\psi_2\psi_2) = 1/3$, $(\psi_1, \psi_2) = (\psi_2, \psi_1) = 1/2$].

Операции (8.21) выполняются разделяющим устройством, изображенным на рис. 8.8. В отличие от устройств разделения ортогональных сигналов, здесь добавляется устройство формирования весовых функций (η_1 и η_2), которое из функций ψ_1 и ψ_2 образует линейные комбинации вида (8.20).

При выборе сигналов для систем многоканальной связи с разделением по форме часто используется ортогонализация сигналов — операция преобразова-

ния линейно независимых сигналов в ортогональные.

Если, например, ортогонализовать совокупность линейно независимых функций

$$\psi_0 = 1, \quad \psi_1 = t, \quad \psi_2 = t^2, \quad \ldots$$

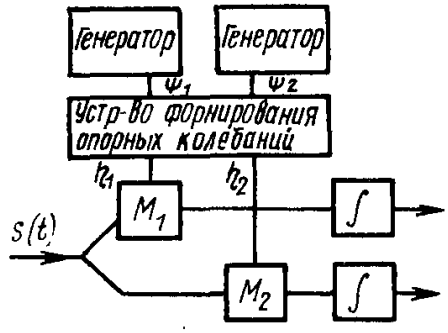


Рис. 8.8. Структурная схема разделения линейно независимых сигналов

на некотором интервале $a<t<b$, то получим систему ортогональных многочленов. В частности, если интервал $[a, b]$ принять равным $[a=-1, b=1]$, то с точностью до постоянных множителей придем к полиномам Лежандра:

$z_0 = 1;$ $\qquad z_4 = 1/8\,(35t^4 - 30\,t^2 + 3);$

$z_1 = t;$ $\qquad z_5 = 1/8\,(63\,t^5 - 70\,t^3 + 15\,t^2);$

$z_2 = 1/2\,(3t^2 - 1);$ $\qquad z_6 = 1/16\,(231\,t^6 - 315\,t^4 + 105\,t^2 - 5);$

$z_3 = 1/2\,(5t^3 - 3);$ $\qquad \ldots\ldots\ldots\ldots\ldots$

Такие сигналы могут использоваться в системах с разделением сигналов по форме. Здесь канальные сигналы передаются одновременно, их частотные спектры перекрываются. Важно подчеркнуть, что в пределах каждого интервала $(-1 \leqslant t \leqslant 1)$ значения коэффициентов c_k должны быть постоянными, чтобы не нарушать условия ортогональности полиномов.

Совокупность линейно независимых функций (8.18) можно ортогонализовать и другими способами; так мы придем к иным полиномам, например Чебышева, Лаггера, Эрмита и другим. На основе некоторых из них также разрабатываются многоканальные системы передачи информации.

В последние годы успешно развиваются цифровые методы разделения сигналов по их форме, в частности, в качестве переносчиков различных каналов используются дискретные ортогональные последовательности в виде функций Уолша, Радемахера и другие. Функции Радемахера образуются из синусоидальных функций с помощью соотношения: $r_k(\theta) = \operatorname{sign}[\sin(2^k \pi\theta)]$, $0 \leqslant \theta < 1$, где аргумент $\theta = t/T$ — безразмерное время; T — период функции, а положительное целое число $k = 0, 1, 2, \ldots$ — порядок функции; sign x — знак действительного числа x: $\operatorname{sign} x = +1$ при $x > 0$ и $\operatorname{sign} x = -1$ при $x < 0$. Иначе говоря, функции Радемахера, принимающие значения ± 1, можно трактовать как функции «прямоугольного синуса». На рис. 8.9 приведены в качестве примера графики первых четырех функций Радемахера $r_k(\theta)$ для $k = 0, 1, 2, 3$. Легко видеть, что функции $r_k(\theta)$ ортонормированы на интервале $0 \leqslant \theta < 1$:

$$\int_0^1 r_i(\theta)\, r_k(\theta)\, d\theta = \begin{cases} 1 & i = k; \\ 0 & i \neq k. \end{cases}$$

Дальнейшим развитием системы функций, имеющих форму «прямоугольной волны», является система функций Уолша $\{\text{wal}(m, \theta)\}$. Она образуется следующим образом. По определению вводится функция $\text{wal}(0, \theta) = 1$ при $m = 0$. Для получения функции $\text{wal}(m, \theta)$ при $m \geqslant 1$ достаточно записать число m в двоичной системе счисления, т. е. представить суммой

$$m = 2^{\mu_1} + 2^{\mu_2} + \ldots + 2^{\mu_p},$$

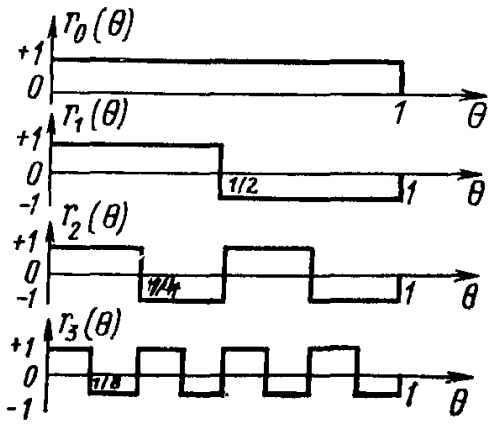


Рис. 8.9. Графики первых четырех функций Радемахера

где $\mu_1 < \mu_2 < \ldots < \mu_p$ — положительные целые числа. При этом функция Уолша определяется формулой

$$\text{wal}\,(m,\ \theta) = r_{\mu_1+1}(\theta)\, r_{\mu_2+1}(\theta) \ldots r_{\mu_p+1}(\theta).$$

Порядок k функции Уолша $\text{wal}_k(\theta)$ равен числу знакоперемен на интервале $[0,\ 1/2]$ и определяется как $k = \frac{m}{2}$ для четных m и $k = (m+1)/2$ для нечетных m. На рис. 8.10 приведены графики первых восьми функций Уолша wal (0, θ), wal (1, θ), ..., wal (7, θ), построенных по четырем функциям Радемахера.

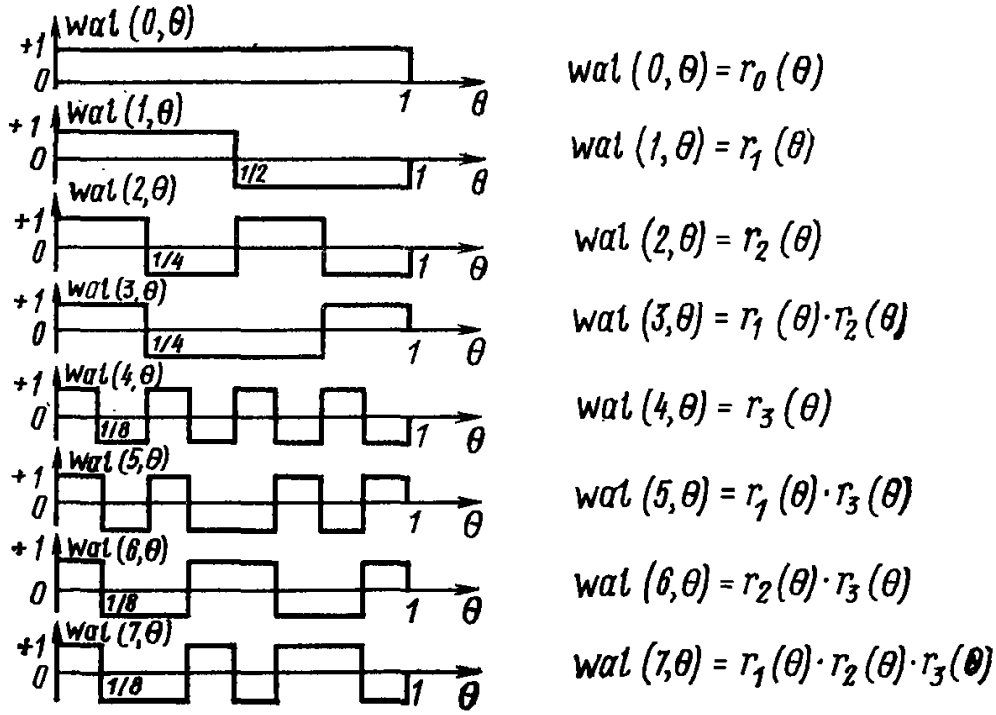


Рис. 8.10. Графики первых восьми функций Уолша

Легко убедиться в том, что функции Уолша не только ортогональны, они обладают свойством мультипликативности. Это означает, что произведение любых двух функций Уолша также является функцией Уолша:

$$\text{wal}\,(i,\ \theta)\ \text{wal}\,(k,\ \theta) = \text{wal}\,(p,\ \theta),$$

где $p = i \oplus k$. В связи с возможностью применения к функциям Уолша логических операций, они являются весьма перспективными при разработке многоканальных цифровых систем передачи с уплотнением по форме с использованием микросхемотехники.

8.4. СПОСОБЫ РАЗДЕЛЕНИЯ СИГНАЛОВ В АСИНХРОННЫХ АДРЕСНЫХ СИСТЕМАХ СВЯЗИ

Рассмотренные в предыдущих параграфах системы многоканальной связи с ортогональными сигналами требуют для нормальной работы той или иной синхронизации:

точного совпадения спектра сигнала с полосой пропускания канала при частотном разделении;

точного совпадения временны́х интервалов передачи сигналов отдельных каналов при временно́м разделении;

точного определения моментов начала и конца тактового ин-

тервала в системах с разделением сигналов по форме активными фильтрами;

точной установки момента отсчета в системе с согласованными фильтрами.

В ряде случаев осуществить точную синхронизацию затруднительно. С подобными ситуациями приходится сталкиваться, например, при организации оперативной связи большого числа подвижных объектов (самолетов, автомобилей). Такая задача возникает при организации оперативной связи с использованием искусственных спутников Земли в качестве ретрансляторов. Во всех этих случаях становятся предпочтительными системы асинхронной многоканальной связи, когда сигналы всех абонентов передаются в общей полосе частот, а каналы не синхронизированы между собой. Поскольку в таких системах за каналами не закреплены ни частотные, ни временны́е интервалы и время работы каждого канала произвольно, то такие системы называют системами *со свободным доступом* к *линии связи* или системами с незакрепленными каналами. В системах со свободным доступом каждому каналу (абоне́нту) присваивается определенная форма сигнала, которая и является отличительным признаком («адресом») данного абонента. В отличие от обычного разделения по форме, где условие ортогональности сигналов выполняется лишь тогда, когда тактовые интервалы всех каналов жестко синхронизированы, для возможности полного линейного разделения сигналов в системе со свободным доступом ортогональность (или, по крайней мере, линейная независимость) между любой парой сигналов должна сохраняться при любом сдвиге по времени. Это значит, что для пары сигналов, $s_i(t)$ и $s_k(t)$, должно выполняться условие

$$(s_i(t),\ s_k(t+\tau)) \equiv \int_T s_i(t)\, s_k(t+\tau)\, d\tau = 0 \qquad (8.22)$$

при $0<\tau<T$, где T — длительность элемента сигнала, а интегрирование производится на любом интервале времени длительностью T.

Можно показать, что условие (8.22) выполняется только в случае, когда сигналы $s_k(t)$ представляют собой белый шум, т. е. имеют неограниченную ширину спектра и бесконечную дисперсию. Для реальных же сигналов условие (8.22) невыполнимо. Можно, однако, построить сигналы, для которых (8.22) выполняется приблизительно, в том смысле, что

$$(s_i(t),\ s_k(t+\tau)) \ll \| s_i(\widetilde{t}) \|^2 \quad \text{при} \quad 0<\tau<T, \qquad (8.23)$$

т. е. скалярные произведения сигналов при любом сдвиге по времени много меньше энергии элемента сигнала. Такие системы сигналов можно назвать «почти ортогональными». Из предыдущего нетрудно догадаться, что такие сигналы должны иметь очень широкий спектр, или, точнее, большую базу (произведение FT). Наиболее распространенным примером технического использования почти ортогональных сигналов могут служить определенным об-

разом подобранные последовательности дискретных, в частности двоичных, импульсов. Каждому каналу присваивается одна из множества почти ортогональных и, следовательно, различимых дискретных последовательностей, которая является «адресом» канала. Это приводит к названию «асинхронные адресные системы связи» (ААСС) или «дискретно-адресные системы связи». Использование почти ортогональных сигналов является дальнейшим развитием способов разделения сигналов по форме.

Важным достоинством ААСС является то, что здесь нет необходимости в центральной коммутационной станции; все абоненты имеют прямой доступ друг к другу без частотной перестройки приемных и передающих устройств (рис. 8.11). Подобно системам автоматической телефонной связи (АТС) здесь достаточно «набрать адрес» вызываемого абонента, т. е. изменить «форму» последовательности импульсов, или, что то же, изменить закон чередования импульсов в последовательности.

В системах с закрепленными каналами добавление хотя бы одного нового абонента оказывается возможным лишь при исключении одного из имеющихся в системе. Значительно проще эта задача решается в системах ААСС. Здесь, вследствие свободного доступа к линии связи, могут вести передачу любые N_0 абонентов из общего числа N абонентов системы связи. При определении числа N_0 нужно учитывать, что вследствие неполной ортогональности сигналов в ААСС неминуемы переходные помехи («шумы неортогональности»), уровень которых быстро растет с увеличением N_0. Поэтому число *одновременно* работающих абонентов должно быть ограничено. Допустимое значение N_0 возрастает с увеличением базы сигнала, так как чем больше база, тем точнее выполняются условия (8.23).

В зависимости от степени «активности» абонентов (т. е. от доли времени, занимаемого k-м каналом для передачи сообщений) можно организовать, например, 1000-канальную систему связи, в которой одновременно ведут передачу любые 50 абонентов из 1000. В таких системах легко реализуются резервы пропускной способности, возникающие за счет «мало активных» абонентов. Изучив статистику сообщений, передаваемых по каждому каналу, можно установить допустимое число каналов в системе N, при котором обеспечивается нормальная работа N_0 «активных» каналов.

Поскольку асинхронные адресные системы многоканальной связи имеют ряд интересных в принципиальном отношении особенностей, рассмотрим наиболее важные из них. В качестве примера почти ортогональных сигналов можно указать на множество реализаций квазибелого шума в ограниченной полосе частот.

Основной особенностью сигналов в виде реализаций ограниченного по полосе частот белого шума является то, что их спектр является равномерным в пределах полосы частот F_0, а их автокорреляционная функция имеет δ-образную форму (острый пик).

Напомним (см. § 4.4), что при прохождении сигнала через согласованный с ним фильтр форма сигнала на выходе фильтра сов-

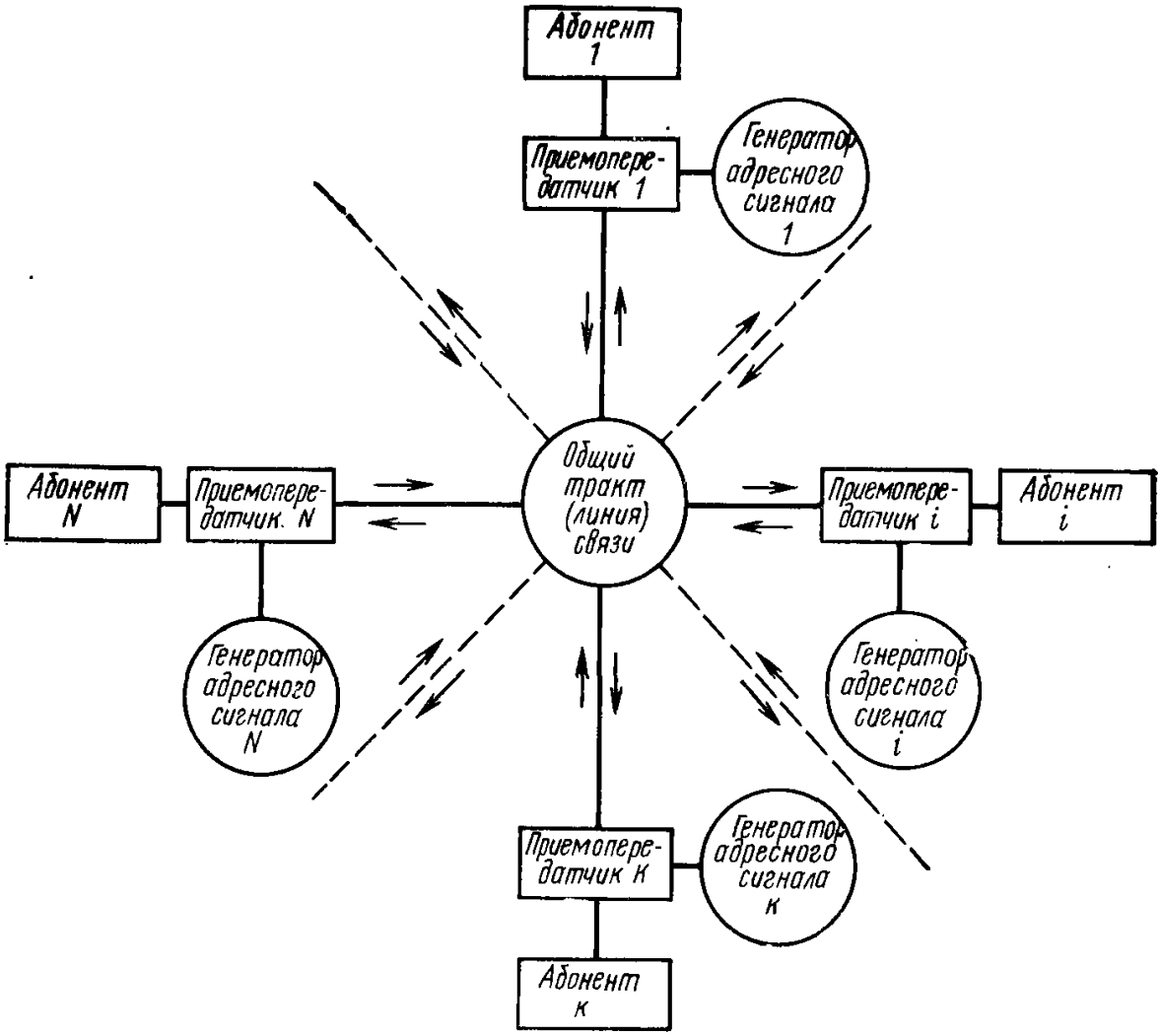


Рис. 8.11. Структура многоканальной асинхронной адресной системы связи

падает с автокорреляционной функцией. В рассматриваемом случае на выходе фильтра, согласованного с «адресом» приходящего сигнала, возникает острый пик напряжения, что облегчает задачу разделения и принятия решения при демодуляции. Этим обусловлена наметившаяся в последние годы тенденция к созданию таких сигналов, чтобы их свойства приближались к свойствам реализаций белого шума, а обработка не вызывала серьезных технических затруднений. Сигналы с такими свойствами называются *шумоподобными.* Функция автокорреляции шумоподобного сигнала должна приближаться по форме к функции автокорреляции белого шума. Вместе с тем, применительно к системам многоканальной связи возникает еще одно существенное требование: взаимная корреляция сигналов должна быть минимально возможной.

Примеры шумоподобных сигналов. В настоящее время усиленно разрабатываются методы синтеза сигналов с заданными автокорреляционными и взаимно корреляционными свойствами. Если рассматривать последовательности из n импульсов прямоугольной формы, которые могут принимать значения ± 1, то простым подбором можно найти такие последовательности,

для которых $B(0)=1$, а $|B_n|=1/n$. Среди них, прежде всего, назовем *последовательности Баркера* (табл. 8.1).

Таблица 8.1

Число импульсов	Номер импульса													Максимум модуля АКФ	
	0	1	2	3	4	5	6	7	8	9	10	11	12	основной	побочные
2	+1	−1												1	1/2
3	+1	+1	−1											1	1/3
4	+1	+1	−1	+1										1	1/4
5	+1	+1	+1	−1	+1									1	1/5
7	+1	+1	+1	−1	−1	+1	−1							1	1/11
11	+1	−1	+1	+1	−1	+1	+1	+1	−1	−1	−1			1	1/11
13	+1	+1	+1	+1	+1	−1	−1	+1	+1	−1	+1	−1	+1	1	1/13

Последовательности Баркера имеют близкую к идеальной форме автокорреляционную функцию; абсолютная величина боковых лепестков не превышает величины $1/n$ главного значения. На рис. 8.12 приведены последовательность Баркера для $n=11$ и ее автокорреляционная функция. Прием последовательности s_1 рис. 8.12*а* («адрес» I канала) выполняется согласованным трансверсальным фильтром рис. 8.13, структура которого сходна с рис. 4.9. Здесь импульсы последовательности Баркера с числом знаков $n=11$ поступают на линию задержки ЛЗ, имеющую отводы через каждые Δt, далее на фазоинверсные (—) и фазосохраняющие (+) каскады с одинаковыми коэффициентами передачи, схему суммирования Σ и решающее устройство РУ. Фазоинверсные и фазосохраняющие каскады включены в порядке, соответствующем чередованию биполярных импульсов n-последовательности; число этих каскадов равно числу знаков последовательности. Первый каскад включен до линии задержки, последний — на конце ЛЗ. При приеме n-последовательность продвигается по ЛЗ, и в

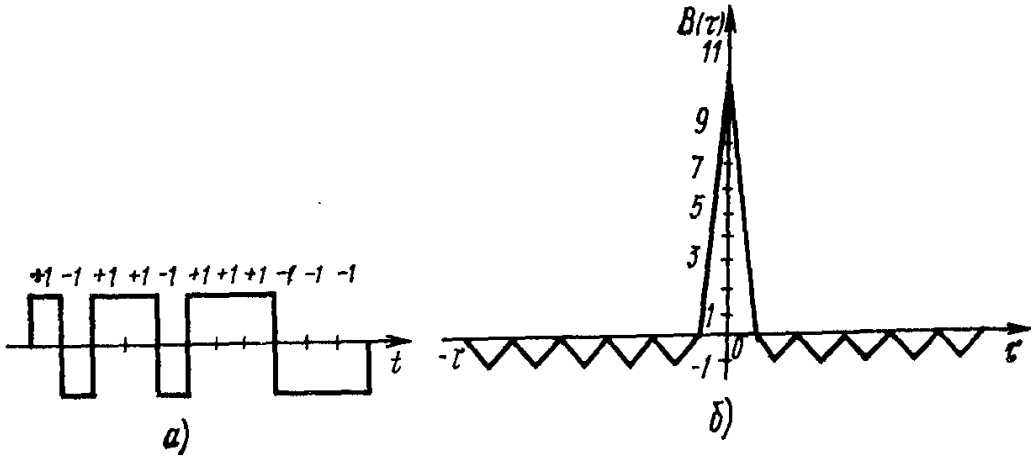


Рис. 8.12. Последовательность Баркера при $n=11$ и ее автокорреляционная функция

момент, когда все импульсы последовательности совпадут по знаку с каскадами, включенными между отводами ЛЗ и суммирующим устройством, все импульсы складываются синфазно, на выходе РУ появляется наибольший импульс — согласованный фильтр зафиксирует «адрес» I канала. При всех других сдвигах суммирование производится не в фазе (с разными знаками), и на входе решающего устройства появляются максимальные уровни, не превышающие по модулю величины $1/n$ максимального значения. Поскольку функция взаимной корреляции между последовательностями также имеет наибольшие значения, не превышающие $1/n$, то последовательность «адреса» II канала также не может вызвать ложного срабатывания решающего устройства I канала.

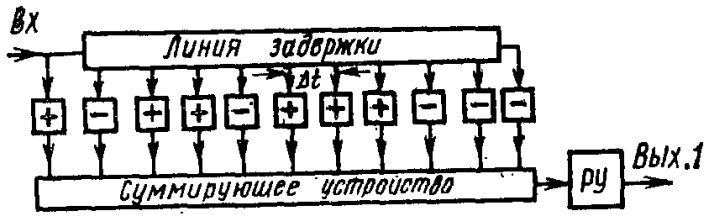


Рис. 8.13. Согласованный фильтр для последовательности Баркера $n=11$

Баркером были указаны последовательности, имеющие «остатки» автокорреляционных функций $|b_k|=1/n$ только для $n\leqslant 13$. Исследования показали, что последовательностей с «остатками» величины $1/n$ для $n>13$ не существует. Поэтому для больших n приходится довольствоваться последовательностями, имеющими «остатки» бо́льшие, чем $1/n$.

Несколько худшие автокорреляционные функции по сравнению с баркеровскими последовательностями, но все же достаточно подходящие для использования в качестве сигналов имеют *линейные рекуррентные* последовательности (ЛРП), или, как их называют иначе, «линейные последовательности сдвигового регистра максимальной длительности». Для ЛРП отношение главного максимума к максимальному боковому лепестку автокорреляционной функции растет приближенно как $\sqrt{n}$, где n — число импульсов в последовательности длительности T.

ЛРП обладают свойством «хаотичности», которое заключается в следующем. Если из периода ЛРП, содержащего $n=2^{\mu}-1$ членов выбрать все возможные отрезки по μ членов в каждом, то, во-первых, среди этих отрезков не будет совпадающих и, во-вторых, среди них найдутся любые комбинации из $+1$, -1, состоящие из μ членов (кроме «запрещенной» комбинации, состоящей только из $+1$). Эти свойства сходны со свойствами случайных биполярных последовательностей; поэтому ЛРП часто называют *псевдослучайными* или шумоподобными последовательностями. К тому же автокорреляционная функция ЛРП имеет форму, сходную с автокорреляционной функцией квази белого шума с ограниченным спектром. Число элементов последовательности n с увеличе-

нием μ резко возрастает, практически удваиваясь при увеличении μ на единицу:

μ...	2	3	4	5	6	7	8	9	10	11	12
n...	3	7	15	31	63	127	255	511	1023	2047	4095

Так как, в принципе, n можно выбрать сколь угодно большим, то автокорреляционная функция может быть получена весьма близкой к идеальной; можно показать также, что среди длинных последовательностей существуют такие, для которых максимумы взаимно корреляционной функции не превышают $1/\sqrt{n}$, т. е. могут быть сделаны достаточно малыми при больших n; то же относится к боковым лепесткам автокорреляционных функций.

ЛРП формируются генераторами двоичных импульсов с использованием сдвигового регистра. Прием ЛРП может осуществляться как согласованными фильтрами, так и корреляторами.

При передаче псевдослучайных последовательностей по высокочастотному каналу обычно применяется фазовая модуляция несущей: например, импульс $+1$ преобразуется в $a\cos\omega_0 t$, а импульс -1 — в $a\cos(\omega_0 t+\pi)$.

Из других способов формирования шумоподобных сигналов для асинхронных адресных систем связи упомянем о способе формирования с помощью частотно-временно́й матрицы (ЧВМ). Мы уже отмечали, что в системах с ортогональными сигналами энергия каждого сигнала полностью отделяется от энергии других сигналов. Это положение становится наиболее отчетливым, если обратиться к частотно-временны́м диаграммам системы связи при частотном разделении (рис. 8.14*а*) и при временно́м разделении сигналов (рис. 8.14*б*). Здесь каждому каналу отводится определенная область частотно-временно́го пространства; положение площадки можно рассматривать как «адрес» абонента. Однако частотно-временну́ю диаграмму можно разделить на площадки иным способом: «адрес» каждого канала можно сформировать из набора «элементарных площадок» частотно-временно́го пространства (рис. 8.15*в*). Здесь каждый информационный символ (рис. 8.15*а*) отображается определенной последовательностью импульсов, имеющих разные частоты (рис. 8.15*б*). Эти «адресные» наборы импульсов составляются на основе их представления в виде частотно-временно́й матрицы (рис. 8.15*в*); к ним предъявляются обычные требования хорошей свертки (с малыми боковыми лепестками) автокорреляционных функций и малой взаимной корреляции. Изменение временно́го положения импульсов и различие в их частотах дают возможность сравнительно простыми техническими средствами получить несколько тысяч частотно-временны́х комбинаций («адресов»). Разумеется, не все комбинации частотно-временно́й матрицы используются в качестве адресов; среди них встречаются и такие, которые не обладают необходимыми корреляционными свойствами. Максимальное число адресов, для которых уровень боковых лепестков корреляционных функций не превышает $1/\sqrt{FT}$, приближается к FT. Сигналы ЧВМ являются раз-

новидностью сигналов, различающихся по форме. Их также можно разделять как согласованными фильтрами, так и корреляторами.

Развитие асинхронных адресных систем во многом будет определяться тем, насколько быстро и успешно разрешится проблема отыскания нужного числа почти ортогональных сигналов с хоро-

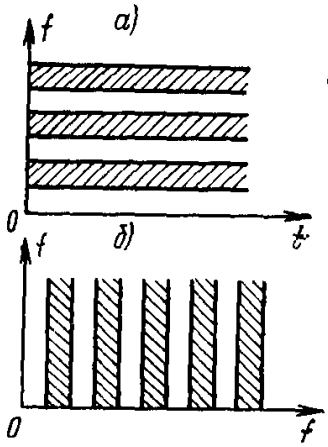


Рис. 8.14. Частотно-временны́е диаграммы при частотной *(а)* и временно́м *(б)* разделении

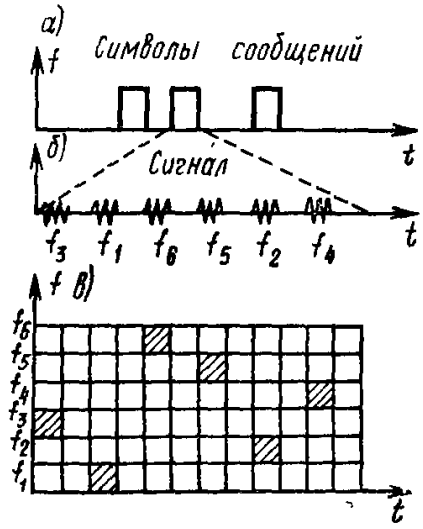


Рис. 8.15. Формирование шумоподобного сигнала с помощью частотно-временно́й матрицы

шими автокорреляционными свойствами и малыми взаимно корреляционными остатками. В последние годы успешно разрабатываются системы, в которых в качестве «адресов» используются наборы смещенных по частоте на величину $\Delta f = 1/T$ псевдослучайных последовательностей ЛРП, т. е. адрес каждого канала представляется в виде сигнала

$$s_k(t) = \exp\left\{i\left[(\omega_0 + k\Delta\omega)t + \frac{\pi}{2}\varphi(t)\right]\right\},$$

где $\varphi(t)$ $(+1, -1, -1, -1, +1, -1, \ldots)$ — псевдошумовая последовательность. В ряде систем используются различные сочетания способов формирования адресов посредством ЧВМ и ЛРП; часто применяются многофазные коды.

Интересной в принципиальном отношении разновидностью способов разделения сигналов в ААСС является один из способов статистического разделения.

Сущность его заключается в том, что в трансверсальном фильтре (см. рис. 8.13) суммирующее устройство заменяется логическим блоком совпадения, на выходе которого возникает импульс тогда и только тогда, когда на все его входы одновременно поступают импульсы одного знака. Такая система используется при формировании сигнала по частотно-временно́й матрице. Однако для боль-

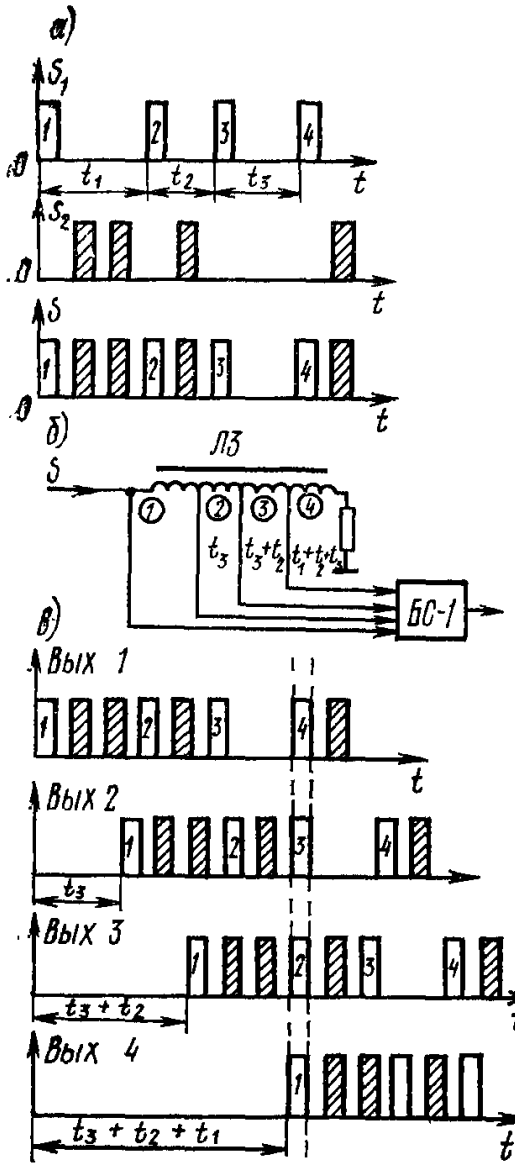


Рис. 8.16. Разделение сигналов в асинхронной адресной системе связи:
а) сигнал на входе разделяющего устройства; *б)* схема разделяющего устройства; *в)* сигналы на выходах линии задержки

шей наглядности мы рассмотрим более простую систему.

Пусть, например, сигнал I канала образован последовательностью из четырех нмпульсов, отстоящих друг от друга на временны́е интервалы t_1, t_2 и t_3; комбинация этих междуимпульсных интервалов образует «адрес» I канала (s_1 на рис. 8.16*а*). Сигнал II канала состоит также из последовательности четырех импульсов; однако интервалы между импульсами последовательности II отличаются от интервалов последовательности I (s_2 на рис. 8.16*а*). Длительности последовательностей I и II могут быть одинаковы и равны, например, T, однако во времени они могут быть смещены друг относительно друга. При передаче по общей линии связи эти последовательности суммируются и на приемной стороне последовательность импульсов имеет вид s на рис. 8.16*а*. Для выделения сигнала I канала («адреса» I абонента) воспользуемся разделяющим устройством, функциональная схема которого изображена на рис. 8.16*б*; схема содержит линию задержки ЛЗ и блок совпадений БС-1. Первый отвод ЛЗ имеет в точке t_3, второй — в точке t_3+t_2, третий — в точке $t_3+t_2+t_1$. При таком выборе отводов ЛЗ совмещает во времени только импульсы «своего» адреса. Иначе говоря, импульс на выходе блока совпадений появится только тогда, когда на его вход поступят одновременно все импульсы (1, 2, 3, 4), принадлежащие сигналу I канала. Импульсы сигнала II канала поступят на блок совпадений I канала в разное время, поэтому на выходе БС-1 импульса не возникнет, т. е. схема рис. 8.14*б* выделит сигнал I канала. Для выделения сигнала II канала потребуется другая ЛЗ с отводами, согласованными с временны́ми интервалами между импульсами последовательности II. Прн одновременной работе многих абонентов из передаваемых ими импульсов может случайно возникнуть последовательность, соответствующая какому-то другому абоненту («ложный адрес»). При достаточно сложных адресах, в частности, образованных частотно-временно́й матрицей, вероятность

такого события невелика. Это позволяет обеспечить асинхронный режим работы значительного числа каналов.

Отметим в заключение, что в технической литературе имеется описание большого числа различных систем асинхронной многоканальной связи. Наиболее характерными из них являются системы на 1000—1500 каналов с 50—100 «активными» абонентами.

8.5. КОМБИНАЦИОННОЕ РАЗДЕЛЕНИЕ СИГНАЛОВ

При передаче дискретных сообщений часто используется комбинационный способ формирования группового сигнала. Сущность этого способа состоит в следующем.

Пусть необходимо организовать передачу N независимых дискретных сообщений по общему групповому каналу. Если каждый элемент сообщения может принимать одно из m_i возможных состояний ($i=1, 2, \ldots, N$), то общее число состояний системы из N источников будет $M=m_1 m_2 \ldots m_N=\prod_{i=1}^{N} m_i$. При одинаковых источниках $m_1=m_2=\ldots=m_N=m$ и, следовательно,

$$M=m^N. \tag{8.24}$$

Таким образом, используя основание кода $M=m^N$, можно передавать одновременно информацию от N индивидуальных каналов, работающих с основанием кода m.

Если, в частности, $m=2$ (элемент сообщения может принимать одно из двух возможных состояний, например «0» и «1»), а число каналов $N=2$, то оказываются возможными четыре разные комбинации элементарных сигналов «0» и «1» в обоих каналах.

Задача теперь сводится к передаче некоторых чисел, определяющих номер комбинации. Эти числа могут передаваться посредством любого кода. При такой передаче групповой сигнал является отображением определенной комбинации сигналов различных каналов. Разделение сигналов, основанное на различии в комбинациях сигналов разных каналов, называется *комбинационным* разделением.

Типичным примером комбинационного разделения является система двукратной частотной модуляции (ДЧМ), иногда называемой двукратным частотным телеграфированием (ДЧТ). Для передачи четырех комбинаций сигналов двух каналов используются четыре разные частоты: f_1, f_2, f_3, f_4; при двукратной фазовой манипуляции (ДФМ) каждой комбинации состояний I и II каналов соответствует определенное значение фазы группового сигнала φ_1, φ_2, φ_3 или φ_4 (табл. 8.2).

В качестве иллюстрации принципа комбинационного разделения рассмотрим пример разделения сигналов при двухканальной системе частотного телеграфирования (рис. 8.17). Здесь принятый сигнал разделяется фильтрами Φ_1, Φ_2, Φ_3, Φ_4, подключенными к детекторам $Д_1$—$Д_4$, попарно работающим на общие нагрузки.

При передаче частоты f_1 напряжение с выхода Φ_1 подводится

через диоды $Д_1$ и $Д_2$ к входным зажимам 0_1 и 0_2 аппаратов I и II каналов. При передаче частоты f_3 напряжение с фильтра $Ф_3$ подключается через диоды $Д_5$ и $Д_6$ соответственно к зажимам 0_1 и 1_2. Все остальные соединения на схеме рис. 8.17 выполнены в соответствии с табл. 8.2.

Таблица 8.2

Канал I	0	1	0	1
Канал II	0	0	1	1
ДЧМ	f_1	f_2	f_3	f_4
ДФМ	φ_1	φ_2	φ_3	φ_4

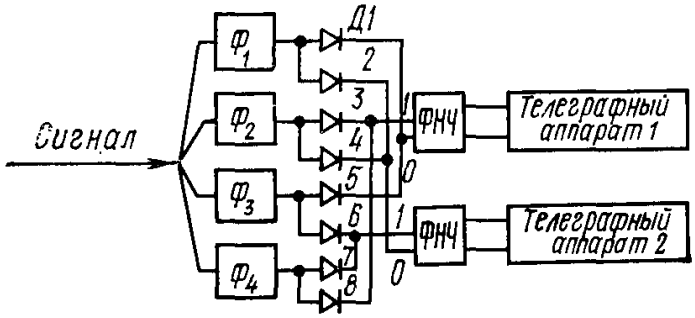


Рис. 3.17. Комбинационное разделение сигнала в системе ДЧМ

При оптимальном приеме для разделения сигналов на частотах f_1, f_2, f_3 и f_4 используются не полосовые, а согласованные фильтры. Если частотные интервалы между f_i и f_k удовлетворяют условию ортогональности, то вероятность ошибки p в одном из каналов ДЧМ при оптимальном некогерентном приеме определяется так:

$$p = e^{-\frac{1}{2}h^2} - \frac{2}{3}e^{-\frac{2}{3}h^2} + \frac{1}{6}e^{-\frac{3}{4}h^2}. \qquad (8.25)$$

Сравнение системы ДЧМ с обычной двухканальной ЧМ системой с частотным разделением показывает, что обе системы занимают практически одинаковую полосу частот, однако мощность сигнала, требуемая для обеспечения заданной верности, при ДЧМ почти вдвое меньше, чем при частотном уплотнении. Существенно меньше оказывается и пиковая мощность при ДЧМ. Поэтому в системах с ограниченной энергетикой комбинационное разделение по методу ДЧМ находит широкое применение.

Комбинационные ДФМ системы на практике реализуются в виде двойной относительной фазовой модуляции ДОФМ по тем же причинам, по которым вместо абсолютных систем ФМ используются относительные — ОФМ. Аналогично можно строить системы комбинационного уплотнения для большего числа каналов — многократную частотную (МЧМ), многократную относительную фазовую модуляцию (МОФМ) и др.

В случае МЧМ, при выборе частот, обеспечивающих ортогональность системы передаваемых сигналов, занимаемая полоса частот с ростом N увеличивается также экспоненциально. Вероятность ошибки в каждом канале с увеличением N также возрастает, но очень медленно. Поэтому такие системы применяют в тех случаях, когда используемый канал связи обладает большими частотными ресурсами, но энергетические его возможности ограничены.

В случае МОФМ, наоборот, занимаемая полоса частот с ростом N почти не расширяется, но вероятность ошибки увеличивается очень быстро и для сохранения требуемой верности необходимо увеличивать мощность сигнала. Такие системы пригодны в тех ситуациях, когда существуют жесткие ограничения полосы пропускания канала, а мощность сигнала практически не лимитирована.

Подробно многоканальные системы связи изучаются в специальных технических курсах.

ЗАКЛЮЧЕНИЕ

Фундаментальными работами В. А. Котельникова и К. Шеннона было положено начало современной теории передачи сообщений. В трудах советских и зарубежных ученых эта теория получила дальнейшее развитие. Классическая теория помехоустойчивости при флуктуационных помехах развита для каналов со случайно изменяющимися параметрами и продолжает развиваться в направлении учета реальных характеристик сигналов и помех, в том числе нестационарных. Вопросы синтеза оптимальных приемников непрерывных и импульсных сигналов успешно решаются на основании теории нелинейной фильтрации. Дальнейшим шагом является разработка и применение методов построения инвариантных и адаптивных систем, позволяющих обеспечить высокую достоверность передачи сообщений в каналах с переменными параметрами при неполной априорной информации о сигналах и помехах.

Примером сложной экономически обоснованной системы должна стать *Единая автоматизированная система связи* нашей страны /ЕАСС/.

В существующей сети связи передача информации ведется посредством телефона, телеграфа, фототелеграфа, телевидения. Передача этих столь различных сообщений предъявляет совершенно различные требования к оконечной аппаратуре, к линиям и коммутационным устройствам. Имеющиеся линии различных видов связи имеют существенно различные характеристики и эксплуатационные свойства. Все это создает большие трудности при создании единой системы связи.

ЕАСС будет представлять собой единую в организационном и техническом отношении систему, позволяющую передавать любую информацию между любыми пунктами страны. Основными потребителями информации в ближайшем будущем будут вычислительные и управляющие центры. Эта информация будет в основном цифровой. Такой вид информации является универсальным, позволяющим передавать как дискретные, так и непрерывные сообщения. Следовательно, на базе цифрового способа передачи можно унифицировать передачу любых сообщений и тем самым обеспечить техническое единство ЕАСС.

В единый комплекс будут объединены кабельные, волноводные, световодные, радиорелейные, космические и все другие линии связи.

ЕАСС будет представлять собой крупнейший инженерный комплекс, который должен вобрать в себя всю существующую сеть связи и развиваться путем планомерного ее наращивания в органическом взаимодействии с системой вычислительных и управляющих центров.

Большую роль в ЕАСС будут занимать системы связи через искусственные спутники Земли. Уже на данном этапе развития эти системы привлекают внимание своей экономичностью, несмотря на сравнительно высокую стоимость запуска спутников. Через спутники можно передавать любую информацию на огромные расстояния.

Быстрое развитие политической, экономической и культурной жизни всех стран мира вызвало значительное увеличение обмена информацией между странами и континентами. Появилась потребность в системах глобальной связи, т. е. связи в масштабах всего земного шара. Применение спутниковых систем позволяет успешно решать и эту задачу.

СПИСОК ЛИТЕРАТУРЫ

1. **Андреев В. С.** Теория нелинейных электрических цепей. М.: Связь, 1972.
2. **Белецкий А. Ф.** Основы теории линейных электрических цепей. М.: Связь, 1967.
3. **Вентцель Е. С.** Теория вероятиостей. М.: Наука, 1964.
4. **Возенкрафт Дж., Джекобс И.** Теоретические основы техники связи. М.: Мир, 1969.
5. **Зюко А. Г., Коробов Ю. Ф.** Теория передачи сигналов. М.: Связь, 1972.
6. **Зюко А. Г.** Помехоустойчивость и эффективность систем связи. М.: Связь, 1972.
7. **Кловский Д. Д.** Теория передачи сигналов. М.: Связь, 1973.
8. **Кловский Д. Д.** Передача дискретных сообщений по радиоканалам. М.: Связь, 1969.
9. **Кловский Д. Д., Шилкин В. А.** Теория передачи сигналов в задачах. М.: Связь, 1978.
10. **Котельников В. А.** Теория потенциальной помехоустойчивости. М.: Энергии, 1956.
11. **Кушнир В. Ф., Ферсман Б. А.** Теория нелинейных электрических цепей. М.: Связь, 1974.
12. **Левин Б. Р.** Теоретические основы статической радиотехники. М.: Сов. радио, т. I, 1974; т. II, 1975; т. III, 1976.
13. **Назаров М. В., Кувшинов Б. И., Попов О. В.** Теория передачи сигналов. М.: Связь, 1970.
14. **Теплов Н. Л.** Теория передачи сигналов. М.: Военное изд-во, 1976.
15. **Тихонов В. И.** Статистическая радиотехника. М.: Сов. радио, 1966.
16. **Тихонов В. И., Кульман Н. К.** Нелинейная фильтрация и квазикогерентный прием. М.: Сов. радио, 1975.
17. **Финк Л. М.** Теория передачи дискретных сообщений. М.: Сов. радио, 1970.
18. **Финк Л. М.** Сигналы, помехи, ошибки. М.: Связь, 1978.
19. **Харкевич А. А.** Избранные труды. Т. III. Теория информации. Опознание образов. М.: Наука, 1972.
20. **Шеннон К.** Математическая теория связи. — В ки.: Работы по теории информации и кибернетике. М.: ИЛ, 1963.
21. **Яглом А. М., Яглом И. М.** Вероятность и информация. М.: Наука, 1973.

СОДЕРЖАНИЕ

Стр.

Основные обозначения 3
Предисловие 5

Глава 1. Системы связи и способы передачи сообщений

1.1. Сообщение и сигнал 7
1.2. Система связи. Канал связи 11
1.3. Помехи и искажения в канале 15
1.4. Кодирование и модуляция 18
1.5. Демодуляция и декодирование 23
1.6. Дискретизация и кодирование непрерывных сообщений 25
1.7. Основные характеристики системы связи 25

Глава 2. Сообщения, сигналы, помехи

2.1. Сообщения, сигналы и помехи как случайные процессы . . . 28
2.2. Спектры случайных процессов 38
2.3. Комплексное и квазигармоническое представление сигналов . . 47
2.4. Огибающая и фаза случайного процесса 51
2.5. Пространство сигналов 57
2.6. Разложение сигналов в ряд по ортогональным функциям . . . 61
2.7. Теорема Котельникова 66
2.8. Информационные параметры сообщений и сигналов 70
2.9. Взаимная информации 76
2.10. Эффективное кодирование дискретных сообщений 79
2.11. Информация в непрерывных сигналах 80

Глава 3. Каналы связи и их пропускная способность

3.1. Общие сведения о каналах связи 84
3.2. Прохождение сигналов через каналы с детерминированными характеристиками 86
3.3. Случайные преобразования сигналов в каналах связи 92
3.4. Математические модели каналов связи 97
3.5. Уравнения состояния. Марковские модели каналов связи . . . 104
3.6. Пропускная способность канала связи 107
3.7. Теорема кодирования для канала с помехами 112

Глава 4. Передача дискретных сообщений в непрерывных каналах

4.1. Прием сигналов как статистическая задача 117
4.2. Критерии качества приема дискретных сообщений 120
4.3. Оптимальные алгоритмы приема при полностью известных сигналах (когерентный прием) 124
4.4. Реализация алгоритма оптимального приема на основе согласованных фильтров 131
4.5. Потенциальная помехоустойчивость при точно известном ансамбле сигналов 139
4.6. Оптимальный прием при сигналах с неопределенной фазой (некогерентный прием) 146
4.7. Прием дискретных сообщений в каналах с замираниями. Разнесенный прием 158
4.8. Прием дискретных сообщений в каналах с сосредоточенными и импульсными помехами 165

Глава 5. Основы теории кодирования

5.1. Назначение и классификация кодов 168
5.2. Принципы помехоустойчивого кодирования 172
5.3. Линейные двоичные блочные коды 179
5.4. Некоторые разновидности систематических кодов 184
5.5. Эквивалентная вероятность ошибки 189
5.6. Системы с обратной связью 190

Стр.

Глава 6. Теория передачи непрерывных сообщений

6.1. Источник непрерывных сообщений и его производительность. Верность передачи непрерывных сообщений 194
6.2. Непосредственная передача сообщений. Оптимальная линейная фильтрация 199
6.3. Оптимальная оценка непрерывных параметров сигнала 207
6.4. Нормальные и аномальные ошибки 211
6.5. Оптимальный прием непрерывных сообщений 216
6.6. Помехоустойчивость систем аналоговой передачи при слабых помехах 223
6.7. Порог помехоустойчивости 228
6.8. Помехоустойчивость систем с поднесущими 233
6.9. Помехоустойчивость систем с импульсной модуляцией 234
6.10. Теория нелинейной фильтрации 236

Глава 7. Цифровые методы передачи непрерывных сообщений

7.1. Общие сведения о цифровой передаче непрерывных сообщений . . 241
7.2. Помехоустойчивость импульсно-кодовой модуляции 244
7.3. Кодирование с предсказанием 251
7.4. Эффективность систем связи 255

Глава 8. Теория многоканальной передачи сообщений

8.1. Основы теории разделения сигналов 260
8.2. Частотное, временное и фазовое разделения сигналов 265
8.3. Разделение сигналов по форме 272
8.4. Способы разделения сигналов в асинхронных адресных системах связи 274
8.5. Комбинационное разделение сигналов 283
Заключение 285
Список литературы 286

Теория передачи сигналов